Data mining and data integration in biology

Ólason, Páll Ísólfur; Brunak, Søren

Publication date:
2008

Citation (APA):
Ólason, P. Í., & Brunak, S. (2008). Data mining and data integration in biology.

Fyrir Helenu Evu og Ólöfu Söru

Foreword

This thesis is the result of my stay at the Center for Biological Sequence Analysis, BioCentrum at the Technical University of Denmark. My Ph.D. work was made possible by grants from the BioSapiens Network of Excellence, FP6 and the Danish Platform for Integrative Biology with Focus on Systemic Proteomics - DPIP. The work was carried out under the supervision of Professor Søren Brunak with the help of a number of people.

Páll Ísólfur Ólason, Reykjavík, March 2007

Acknowledgments

They say that Ph.D. work is a test of one's abilities to work independently, but I would never have finished without the help of a number of people. To name a few of them:

My supervisor: Professor Søren Brunak. Many thanks to him for giving me this opportunity. Always full of novel ideas and inspiring his troops to think in new ways, he deserves a lot of credit.

BioSapiens NoE, especially my DAS and ENCODE collaborators. The work associated with this huge proteomics consortium has been both successful and enjoyable.

Funding agencies. The Danish Platform for Integrative Biology, DPIB, and the Danish National Research Foundation. Thanks for your generousity.

My officemates Olga Rigina and Hans-Henrik Stærfeldt. I felt a bit like James Bond when told that I would be sharing an office with a mathematician from Siberia and a super-hacker. Olga and Hans-Henrik, thank you for all your support.

Co-authors of my papers all 50+ of them, look inside for names.

The CBS administrators. Lone, Johanne, Anne, Dorthe, Marlene, Hanne. You keep CBS running, thanks for all your help.

The system and database administrators. Kristoffer Rapacki, Hans-Henrik Stærfeldt, Peter Wad, Olga Rigina. Thanks for keeping things running and for all your technical wisdom.

The disease gene finding group. Exciting project, great people!

Thomas Sicheritz-Pontén. My collaborator on protein-protein interaction prediction, a sometimes depressing and ultimately unrewarding project. Thomas was also the only person to proofread this thesis, but he will get a reward for that. Thanks Thomas!

The Integrative Systems Biology group. The coolest group, by far ;)

Comrades made through travels, coursework, meetings and projects. My stay at CBS has allowed me to meet and interact with so many great people I don't know where to start. Kasper, Olof, Ulrik, Thomas Skøt, Thomas Pontén, Rodrigo, Zenia, Peter Hallin to name just a few. I know there are many more but it would take pages to put everyone here. Thanks for all the good times.

Kristoffer Rapacki. The official welcoming committee at CBS, Kristoffer strives to make everybody feel at home and always has time to help, despite being extremely busy. Thanks for all your help, humour and friendship.

CBS is a wonderful place to work. Thanks to all you people who make it so.

Open source software authors. This work and this thesis would not have been possible without the generousity of the open source software community.

My family. My lovely wife Ólöf Sara and my daughter, Helena Eva, my greatest source of inspiration, she always helps me put things in perspective. They gave me the energy to finish writing this thesis while working full time on a new job and refurbishing our new apartment in Iceland.

Abstract

Last decade saw an explosion in DNA sequencing and the draft version of the human genome. Now, proteomics is experiencing the same growth. With proteins being the functional elements of living cells, high-throughput proteomics promises more understanding of cellular functions and the interactions between molecules, the essence of systems biology.

Internet technologies are very important in this respect as bioinformatics labs around the world generate staggering amounts of novel annotations, increasing the importance of on-line processing and distributed systems.

One of the most important new data types in proteomics is protein-protein interactions. Interactions between the functional elements in the cell are a natural place to start when integrating protein annotations with the aim of gaining a systems view of the cell. Interaction data, however, are notoriously biased, erroneous and incomplete. They also necessitate new ways of data preparation as established methods for sequence sets are often useless when dealing with sets of sequence pairs. Therefore careful analysis on the sequence level as well as the integrated network level is needed to benchmark these data prior to use.

The networks, which emerge when interaction data are integrated, form a skeleton to which we can attach other annotation types. Then, using graph theoretical methods, we can identify network structures and infer annotations across the links of physical interactions, thus defining novel functional modules, or in the case of dysfunction: disease modules and genes.

Dansk Resumé

I den såkaldte “post-genomic” æra vi oplever nu, er den humane genomsekvens kendt og sekvenseringsprisen falder hvert år. Nu står kapløbet om at finde sammenhængen mellem de forskellige sekvensprodukter, systembiologiens største mål.

Proteiner er organismers byggesten, og deres egenskaber, funktion og vekselvirkninger er meget vigtige i denne sammenhæng. Integration af protein annoteringer fra hele verden er nødvendig for udvikling inden for systembiologi. Projekter ligesom ENCODE arbejder på koordinering af denne slags annoteringer og analyse af de forskellige produkter af de samme gener.

Internetteknologier er meget vigtige i denne sammenhæng, da bioinformatik og sek vensanalyse fremstiller utrolige datamængder. Dataproduktionen overstiger faktisk Moores lov om vækst af dataprocesseringskapacitet, hvilket gør at distribuerede systemer og real-time processering er at foretrække frem for statiske databaser i mange tilfælde.

En af de vigtige nye datatyper inden for cellebiologi er protein-protein vekselvirkninger. Disse data beskriver proteiners samspil og komplekse funktion i cellen og er derfor en naturlig byggesten for systembiologien. Disse data er dog fulde af støj og meget inkomplette. Derfor er det vigtigt at udvikle metoder til op rensning af data og forudsigelse af ukendte vekselvirkninger, hvilket dog ikke er lige til, da vekselvirkningsdata har protein-par som grundenhed, hvilket gør alle algoritmer mere komplekse og opbevaring samt analyse langt besværligere.

De netværk, der opstår ved integration af vekselvirkningsdata, danner et skelet for de funktionelle elementer, komplekser og pathways i cellen. Grafteoretisk analyse af protein netværk er et vigtigt værktøj til forståelse af cellen som et system. Dette kan indebære teoretiske overvejelser omkring tæthed og konnektivitet af proteiner, samt praktiske forbindelser mellem sygdomsgener der optræder sammen i kompleks eller pathway. Integration af data, der beskriver funktion, phenotype og genetiske fejl, kortlagt på dette skelet af vekselvirkningsdata kan hjælpe os med at prioritere kandidatgener for sygdomme, hvilket kan føre til nye lægemidler i bekæmpelsen af medfødte sygdomme.

Contents

Chapter 1

Introduction

1.1 From genomics to functional genomics to systems biology, but not back again

As I attended an "Introduction to Bioinformatics" course in 2001, two articles announced that the sequencing of the human genome had been finished. So why on earth did I choose this field of research and occupation, when the blueprints to life had already been uncovered?

Well, uncovered does not mean understood. As one evil mastermind[1] once said to 007: "The key to a great story is not *who*, or *what*, or *when*, but ***why***". Our DNA blueprints differ only to an amount of a tenth of a percent, all humans have the remaining 99.9% in common. How then do we explain our different appearances? I and my 4 siblings (3 brothers, 1 sister) were all born with blond hair, why is it that only I and one of my brothers now have dark hair, while the others still have blond hair, and what triggered the change in hair color? Why is it that only I am starting to get gray hairs? Actually, I think this thesis holds the answer to that one, although not literally.

Even though a draft version of the human genome has existed for several years now, that has not meant the end of sequencing. On the contrary; we are producing sequences at an unprecedented rate. The human genome sequence is still incomplete and the assembly of the shorter sequence fragments into larger ones is tricky, so huge chunks of DNA are often shuffled around between builds of the genome. Apart from humans, other organisms are continuously being sequenced, with over 30 (partially or fully) sequenced eukaryote genomes available in the Ensembl browser (`www.ensembl.org`) at the time of writing, and the number of sequenced bacteria reaching into the hundreds. Excitingly, only a few days ago it was reported that Craig Venter - one of the people behind the privately funded version of the human genome [274] - and colleagues have produced an astonishing 7.7 million sequence reads from ocean samples collected around the world, most of them novel sequences. These sequences may assist in as different fields as antibiotics production and the search for new, sustainable energy sources [226]. Sequencing is being performed on a gigantic scale, and large scale proteomics analysis has only just begun. This is only the primary sequence data. Such data have now been produced for 30 years and polymerase chain reaction (PCR) and automatic sequencing make this task a breeze. It is the functional analysis of these sequences takes most of our time.

[1] Elliot Carver (Jonathan Pryce) in *Tomorrow never dies* (1997).

For every new stretch of sequence deposited in the data banks, one can anticipate that hundreds of analysis reports are made; sequence similarity reports, predictions of genes, introns, exons, promoter regions, the list goes on. And this is only for the primary DNA sequence. The central dogma of flow of hereditary information is *DNA to RNA to protein*, and this is not a one to one relationship. A single gene can yield a number of splice variants and when you throw in protein post-translational modifications, native and chaperone-assisted folding, domain and motif definitions, expression, physical and chemical properties, the possibilities for analysis are endless.

Figuring out *how* these individual components function and interact is the key to the great story, the *why*. This thesis is an attempt to describe a number of *hows* and then assemble those to answer a couple of *whys*. But don't worry, there are not that many answers in here, and no suggestions about the meaning of life.

The big *why* is to understand the biology of organisms and cells, why these complex systems have evolved the way they have, and for practical purposes: to understand why things sometimes do not function as they should. The key to this understanding is to figure out *how* these systems, and their subparts work. To integrate all the data on the individual components of the cell and and put them in a cellular, physiological, and evolutionary perspective[2].

Combining all these small *hows* into larger *hows* and, eventually, *whys* is pretty tricky and this is where bioinformatics comes into play. Integrating data in biology is not a mere punch on the stapler. The, sometimes ridiculous, amounts of data being generated are well beyond any simple staple to penetrate. And even if they were not, the diversity of the data is so great that simply collecting and stacking data is not meaningful. One has to churn through the endless flow and try to sort, shuffle, filter and index the relevant stuff. Hopefully one has something of interest to report at the end.

[2] *When* is a relevant question to ask, but at what timescale? If we are talking life in general, then evolution certainly is a huge and important field within bioinformatics and biology. Looking at shorter time spans, the cell cycle, development, apoptosis, to name a few, are events with a temporal facet, although one may argue that this aspect only reflects the *how* and the *why*. The only *when* to be addressed in this thesis is the deadline for its submission.

1.2 The problem: which bread crumbs to follow?

Imagine that you are employed at a record store and you unpack a box of compact discs. You need to register the CD in the store's computer system along with a short description and put the CDs in the right rack (pop, rock, country, classical etc.) in the store. How do you proceed? The obvious choice is to simply listen to the CD and describe and classify it accordingly, and this is what the first music shop owners were forced to do. This however is time consuming and requires expensive equipment. As a short cut, many would simply look for similar CDs; has the artist released a CD before? And if so, is it already registered and described in the computer? If we are in luck, the artist already has been described in the system and we can place the new CDs besides his/her earlier work in the shop. If this is the artist's first album, we might look for subtle clues on the CD cover: five young guys standing in line probably indicates a "boy band", big cars with shiny wheels and girls wearing bikinis on the hood is a dead giveaway for hip hop, a swan on a lake suggests Tchaikovsky and so on. Thus, we may integrate data to make our lives easier. The Internet is very valuable in this respect. It is not always trustworthy but usually it gives us the information we need in a few clicks of the mouse.

Sequence analysis has gone through several phases. Ever since the data format, DNA, was described about 50 years ago, we have been seeking to understand these blueprints for life. At first, we had to build equipment to "play" the tracks and build up a database of different sequences. There was no Internet to help us look up artists and new releases. Today, we have indexed so much, that it usually is not necessary to listen to the tracks yourself, it is quicker to just look up the new tracks by similarity in the database and copy/paste its annotations. Sure you get new, novel tracks in now and then and need to examine them manually to be able to classify and register, but this is usually fun and exciting work. Now, the focus is on making the shop more accessible to the customers: you want to cluster the similar categories together, cater for specific needs, allow for new formats of entertainment, to sell re-mixes, posters and t-shirts alongside people's favorite music, to give people the whole picture instead of just plain CDs. Ensembl's BioMart is a good example of such an approach in sequence analysis. Given an organism, you can download, along with the primary DNA sequence, dozens of annotations to attach to the sequences, ranging from gene synonyms to functional classification of proteins. The problem is not lack of data, but rather to figure out what data are relevant to answer your particular question.

The field of bioinformatics has shifted focus from searching to understanding to putting things in perspective; from sequencing to functional analysis to systems biology. Even though systems biology is a relatively new phrase, the concept is not and tasks in systems biology are mostly the same as the tasks in functional genomics, which was a buzzphrase 5 years ago. Performed at a larger scale maybe and with a more global perspective, but at the core, functional analysis of the individual components is still the way to understand the system.

During the initial sequencing era, it was discovered that man, the pinnacle of intelligent life has blueprints as about as complex (as far as count of known genes is concerned) as the nematode *C. elegans*. We are realising that the phenotypes of organisms are the product of more than the bare count of genes, but rather the amount of splice variants and even the number of times the genes are transcribed. Indeed, new results indicate that the quantity in which genes are transcribed and translated is a key to our phenotype, whether this quantitative variation stems from repeats in the DNA itself [218] or from difference in gene expression [245]. Apart from variability in sequence and expression, it is clear that most or all gene products perform their tasks as a part of a larger unit. Protein complexes assemble and disassemble at specific points in time and space [61] and metabolic and signaling pathways are complex chains of reactions and regulation. Figuring out the link between the components in these complex systems, their control and function, then the link between the modules and the phenotype of organisms is the ultimate goal.

Today there is no shortage of information to annotate primary sequences. At the Center for Biological Sequence Analysis (CBS) website alone (`http://www.cbs.dtu.dk`), there are almost 50 tools to analyse primary sequences, either DNA or proteins. The hard part is to decide what information to integrate, to find the relevant data or tools (usually on-line) and finally integrate them in a meaningful way to answer some of the *whys*.

The phenotypes of inherited diseases are caused by (yet) unidentified genes in most cases, and although a single disease gene can be identified in some cases like cystic fibrosis and Huntington's disease, most cases seem to have a much more complex explanation than a mutation in a single gene. Even in cases where disease susceptibility genes are known, the majority of drugs, designed to target those genes (or, rather, their products), fail clinical trials [90]. Usually because of unpredictable reactions with other components in the cell, disrupting the delicate balance that these components display in vivo. When we have learned how those components function and link to each other, we can start modeling the effects of drug leads in cellular systems, and in time, high resolution bioengineering will perhaps enable us to design drugs for each individual, based on their genotype.

So, we have a huge amount of primary sequence data, and for each stretch of sequence we have virtually unlimited annotation sources. Our task is to figure out where to start; to find out which path leads to understanding of the functions and dysfunctions of cells and organisms. The perspective we wish to achieve is systems biology - a map of the forest - but where do we get it?

1.3 Our solution: the interactome as the backbone of systems biology

The recent interest in systems biology has been sparked by the availability of high-throughput data, especially protein-protein interaction (PPI) data. Several proteome-wide interaction studies have been performed and published in the last 5 years [127, 266, 98, 119, 97]. Figure 1.1 shows the annual amount of papers indexed by PubMed containing the phrases "protein interaction" or "systems biology". Publications containing these phrases are appearing at a rate far surpassing the normalised growth of PubMed databases. It is not a coincidence that PPI data and systems biology have come up in the world together. Systems studies are all about connecting components on some level and as proteins are the cell's functional components, physical interactions are a perfect thread to connect proteins in a systematic fashion.

The emerging network of protein-protein interactions, the interactome, makes a ideal scaffold to which we can glue other annotation types in order to build high-level models of reactions, complexes, pathways, and ultimately, the entire cell [126]. The reason that PPI networks make such an excellent scaffold to build integrated analysis upon is that PPIs lie at the heart of most of the cell's functionality. Whether it is replication, transcription, translation, biosynthesis, degradation, signaling, immune response. Proteins are the cell's workhorse, and to understand the cell, we must understand the functions of proteins.

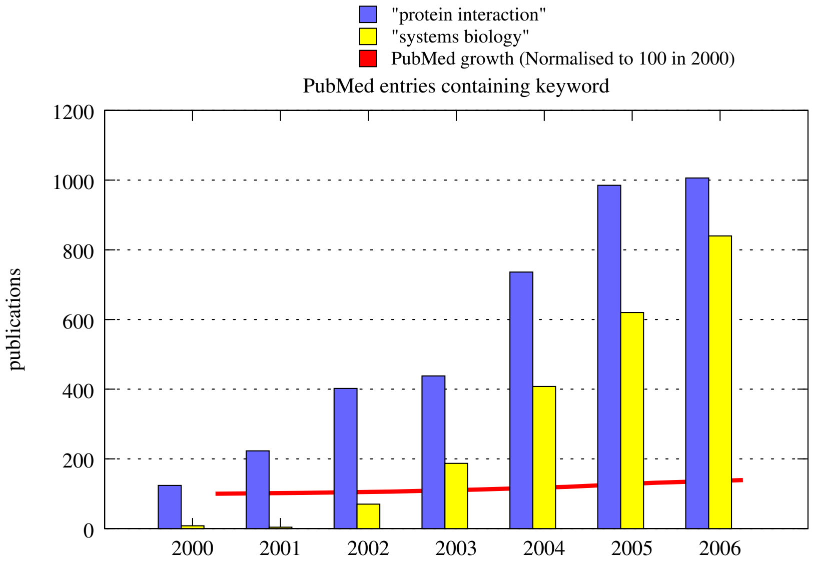


Figure 1.1: **Search results for the phrases "protein interaction" and "systems biology" in the PubMed index. The number of publications containing "protein interaction" has been increasing linearly for the past decade, and following closely, papers mentioning "systems biology" are now appearing at a similar rate. The red line shows the growth of PubMed databases normalised to 100 at the year 2000, included to give a general idea about the publication rate in life sciences. Publication rate of papers containing either of the above mentioned phrases far exceeds normalised PubMed growth.**

PPI data are not perfect, they bring a number of analysis difficulties, as they consist of sets of pairs instead of single proteins. They are also noisy, hard to interpret definitively and incomplete, especially for *H. sapiens* [276].

Having integrated and analysed the interaction data, we want to attach other annotations onto the skeleton of protein interactions. As most annotations come from on-line sources, it is important to know where to look for such data and to be able to trust these. Formatting data using standard specifications is a great help in these situations, easing the workload considerably. These standards need not be technically complex. The most important thing is for everybody to decide on one standard, because what good is having a better standard specification if your lab is the only lab using it?

One of the less frequent data types described later in this text is phenotype data, most of which is on the form of free text, in contrast to most annotation types described in here. There is room and need for much improvement here as systems biology is finally allowing us to use computers to bridge genotype and phenotype in an automated way.

Data integration in biology is still a tough task. Even though chapter 2 describes some advanced ideas and technologies for sharing and integrating data, most data out there needs to be manually glued together for special purpose research. The research papers included in chapters 3 and 4 have all required a huge amount of manual labor to massage the data into results.

1.3.1 Components vs. systems

Having said that the key to understanding the system is to understand the individual components, a systems view of cellular components is quite different from focusing on the details of the individual sequences. Usually a reductionist view is applied, and a perfect understanding of the image on the individual pieces may not be necessary, but we need to make sure they fit together if we are to finish the puzzle.

Molecules in the cell have finite life spans and those molecules (sequences, hormones, nutrients and other) are not the focus of systems biology, their collaborative functions, interactions and regulation is. The components come and go but the system and its subsystems stay the same just like a carpenter still views his hammer as the same he bought 20 years ago, despite having replaced the handle three times and the head once. *Emergent properties* are the properties which arise as data pile up and get integrated, the properties of the system, rather than components. Regardless of emergent properties now being the focus of attention, we still need bottom-up approaches like sequencing and functional analysis of individual sequences because we still do not fully understand these components.

Sequencing is not as expensive or time consuming as it used to be. Proteomics research, on the other hand, is very challenging. The vocabulary of 20 amino acids instead of 4 nucleotides allows for much more complex chemistry. Throw in folding, alternative splicing, post translational modifications, wide dynamic range and tissue, developmental and temporal specificity and we realise that "genes were easy"[3]. Despite not being as scalable and easily automated as DNA sequencing, proteomics is still moving steadily towards high-throughput techniques with "labs on a chip" starting to replace slow, labor-intensive and costly biochemical assays [205, 264].

But experimental techniques are not enough. Large scale data integration that systems biology demands, cannot happen within an isolated lab. With the latest and greatest mass spectrometry equipment and trained staff only being within reach of the best funded labs, data must be made available for colleagues. Guidelines for raw data sharing, like the Bermuda rules [178] for the human genome project, should be adopted for proteomics.

1.4 Directions to the thesis

This thesis is built as an attempt to describe a move from analysis of individual components, proteins in my case, although the concepts of genes and gene products certainly could be (and are) used interchangeably in most cases, to an integrated view of all known components in a system.

In today's post-genomic era, we are certainly moving away from the primary sequence in bioinformatics, the question is how far away? Systems biology was born as a result of high-throughput technologies where thousands of components are analysed at once. The transition to high-throughput methods requires a new way of thinking, new statistics, new algorithms, even new naming schemes, as simply identifying the right sequence in todays monstrous data warehouses is a formidable task. This thesis is about data integration and the structure of the text reflects the workflow.

We start with simple sequences, analysing those with numerous tools in a bottom up fashion; on the desktop, in the lab and most importantly today: on line. We describe the diversity of sequence data and metadata and current work to keep track of it all.

Next, we start putting the protein pieces together, trying to understand how the components interact, and to predict their interactions from primary sequence and domain composition. As we use interactions a skeleton to attach other

[3] The Human Proteome Organization (HUPO) held its inaugural meeting in 2001 under the slogan "Genes were easy".

annotations on, it is important to get this “infrastructure” of systems biology right. PPI data are notoriously noisy and incomplete. Analysis and prediction of PPIs is difficult for many reasons: non-interacting pairs to use as contrast are hard to define, algorithms commonly used for redundancy reduction, similarity searches etc. are designed for sets of single sequences, not sets of sequence pairs.

When the components have been connected, a network emerges and we change our perspective from local to global, using top-down approaches such as network modeling and global trends, such as clustering, to annotate individual components and functional modules, where the individual components work together on a cellular task as a whole. These functional modules, complexes or pathways, sometimes break down, in which case our network analysis of annotations attached to the skeleton of physical interactions may help us identify the faulty component. Identifying the faulty component(s) is the first step to mending a dysfunctional system. Thus we can trace correlated disease phenotypes over links in the network of physical interactions and assign the target proteins as products of potential disease genes.

Chapter 2

Components

In this first chapter of this thesis about biological data integration, I wish to introduce some of the issues in biological data management, with a special focus on protein annotation data. More recent data types - the omics data - are also mentioned. The impact of the Internet on sequence analysis will be discussed, as data integration would be non-existing without publicly available data sets. Finally I will share some thoughts about further developments I envision for data management within on-line biology.

2.1 Biological data

The complexity of molecular biology and our analysis thereof make it hard for us to disseminate the information we have compiled, and to receive, comprehend, and integrate information from others. Two main problems hamper biological data sharing on a global scale:

Data accumulation: The growth rate of biological databases is incredible, its doubling rate surpassing Moore's law [1], which highlights the importance for the field of bioinformatics to keep up the pace.

Data diversity: Biology is a huge field and there is simply an infinite number of data types. To make things even "worse", new data types are continually being made as new experimental designs see the light of day. This makes it hard come up with standardised data sharing formats and infrastructure.

As the world wide web (www) and other technical infrastructures evolve, the first mentioned "problem" of data growth, while still a significant one, must be accepted as a consequence of a productive scientific community and countered with a similar advance in hardware as well as software.

The community's shortcomings in dealing with diversity, however, only become more and more emphasised as we move further towards comparative genomics and systems biology.

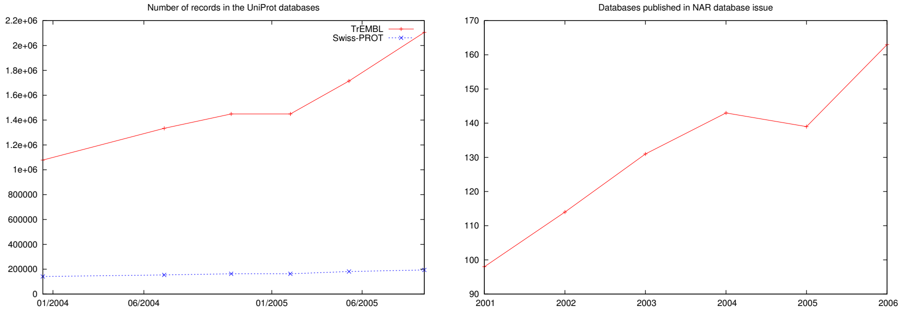


Figure 2.1: **Left: Growth of the UniProt databases: TrEMBL is a database of translated nucleotide sequence entries in EMBL, with automated annotations while Swiss-Prot is a manually curated database with high-confidence annotations. The size is measured as total number of sequences. Right: Each year the *Nucleic Acids Research* magazine publishes a database issue describing new or improved databases pertaining to sequence analysis. The plot shows the number of databases published each year.**

2.1.1 Database growth

The first repositories of sequence data only held a handful of sequences or structures [26, 25], and they were released in paper format or via slow telephone connections if you were so lucky to have a computer in you lab. Today the picture is quite different; as the cost of sequencing drops, sequence databases are overflowing with data and submission rates are increasing. Indeed, the cost of sequencing has fallen more than ten-fold in the last decade, and now plans and funding are available for approaches, that will make an entire human-sized genome sequence available for $1,000 (`http://www.genome.gov/15015208`). Figure 2.1 shows the size of the UniProt databases [288], where TrEMBL is a database of translated nucleotide sequences from EMBL and Swiss-PROT is a manually curated protein database. It is clear that the manual curation effort is lagging about an order of magnitude behind the rate of sequence submissions. This demonstrates the need for high performance computing within biology, as well as high confi-

[1] Moore's law is the empirical observation that the complexity of integrated circuits, with respect to minimum component cost, doubles every 24 months. It is attributed to Gordon E. Moore, a co-founder of Intel [283].

dence, automated annotation strategies. Human effort alone will never keep up with sequence submissions.

The cost of computation rarely grows linearly with the size of input data; rather it grows polynomially or exponentially. In practice, a doubling of the input search database means that the BLAST sequence alignment algorithm [6] needs 8 times the computational time as before the input doubling, because it has n^3 complexity [2] with respect to input data.

Most bioinformatics groups provide some sequence analysis tools and many offer the annotation results as specialised data sets and on-line databases. Therefore there has been an explosion in the number of databases published in the field. *Nucleic Acids Research* publishes an annual database issue which describes new or improved sequence analysis databases. Figure 2.1 shows the steady increase of new databases published in the issue since 2001.

The growing number of databases not only demonstrates the growing amount of data in biological databases, but also the growing number of new data types in biology; next section's subject.

2.1.2 Data types

As shown in the last section, there is no shortage of biological sequence data in the public domain, with databases growing at ever-increasing rates, leaving manual curation efforts far behind. This explosive growth can only be countered by similar performance increases in computing, as well as a continuous effort for more efficient algorithms in bioinformatics.

Following the shift in focus from genomics to functional genomics to systems biology, diversity in data types mirrors the advances in computational biology. Functional genomics has given rise to a huge number of data types, both experimental and computational, used to annotate biological sequences. Twenty years ago, sequence analysis data were mostly limited to the sequences themselves.

[2] "The time complexity of a problem is the number of steps that it takes to solve an instance of the problem as a function of the size of the input (usually measured in bits), using the most efficient algorithm. To understand this intuitively, consider the example of an instance that is n bits long that can be solved in n^2 steps. In this example we say the problem has a time complexity of n^2." "Example: Mowing grass has linear complexity because it takes double the time to mow double the area. However, looking up something in a dictionary has only logarithmic complexity because a double sised dictionary only has to be opened one time more (e.g. exactly in the middle - then the problem is reduced to the half." Taken from [282].

During the last two decades (and even more) the Center for Biological Sequence Analysis (CBS) and other labs have produced sequence annotations, both experimental and computational, for proteins and genes, predicting and analyzing their chemical, structural and functional properties. At CBS' sequence analysis web page alone, `http://www.cbs.dtu.dk/services`, there are currently 9 DNA sequence annotation servers and 31 protein annotation servers; a glimpse at the huge number of different analysis methods (and therefore annotation data types) publicly available.

Moving towards whole genome or proteome analysis, several high-throughput methods have been established. Examples of such new data types in sequence analysis include microarrays [231], DNA chips [172], yeast two-hybrid method [89] and complex purification followed by mass spectrometry [119, 98] for protein protein interaction data, ChIP-chip [219] for protein-DNA binding, protein microarrays [154] etc. Adding to this complexity is a myriad of meta data, which often accompanies lab data such as mass-spectrometry.

Today, the focus is on integrating these bits and pieces of information from functional analysis of sequences; to assemble them into a coherent view.

Omics data

Since the completion of the human genome drafts in 2001 [162, 274] the number of publications containing the phrase "systems biology" has exploded [62]. Along with systems biology come the *omics* data, used to refer to some sort of data in their totality. First there were *genomics* data, closely followed by *proteomics*. Following those came *metabolomics, transcriptomics, nutriomics, interactomics, phenomics, localizomics, spliceomics, ORFeomics* and more. Today, we simply talk about omics data when referring to large scale comparative approaches, where the entities, be it proteins, genes or nutrients, are not the focus, but rather seen as a part of a larger picture.

It is not the aim of this document to survey biological data types exhaustively, as that is an exercise in futility. Furthermore, such a review would soon become obsolete as new data types are constantly emerging. My aim is simply to point at the enormous size of the the problem facing scientists who wish to combine heterogeneous data.

Text mining

The integration of the diverse data types described above is difficult, but the bulk of scientific information lies not in formatted, readily available data sets, but rather in free text in scientific literature.

This short section does not try to give any practical introduction to the field of text mining, which is a whole science in itself, but only to highlight some of the tasks.

The abundance and richness of scientific literature is unfathomable and as I soon learned during my as yet brief research career, it is impossible to stay on top of all the information published, even within a very narrow field such as information exchange standards in biology or protein protein interactions. The MEDLINE biomedical literature database [198] now contains over 15,000,000 searchable abstracts and the PubMed interface at NCBI handles over 80,000,000 queries to the database every month, as seen in figure 2.2.

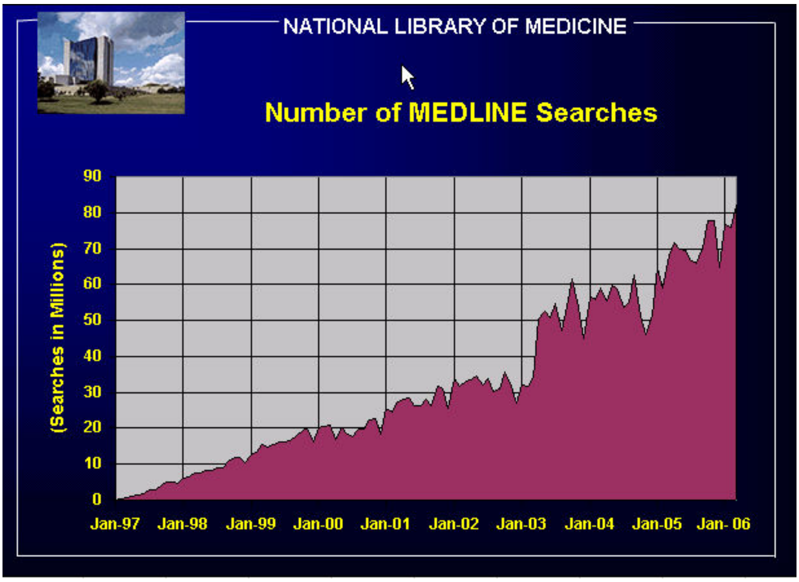


Figure 2.2: **Medline growth. The figure shows the monthly number of queries to Medline. Taken from: http://www.nlm.nih.gov/bsd/medline_growth.HTML.**

Within bioinformatics, a growing effort is made on mining this huge information resource and for the last 5 years a special interest group meeting for text mining in biology, *BioLINK*, has been held at the largest current bioinformatics conference: Intelligent Systems for Molecular Biology (ISMB). CBS has now joined the labs that have incorporated text mining into their arsenal of scientific research tools. CBS' first steps into the field are described in the last chapter (4) of this thesis, where our attempts at candidate gene prioritisation are described.

Literature mining may be performed in various ways and with various tools, ranging from simple string matching and word vector similarity measures to natural language processing and machine learning Some approaches aim at indexing the text for subsequent lookup and extraction, such as a search engine on the Internet might index on line documents. These approaches are termed information retrieval (IR).

A special sub-problem of scientific text mining is to recognise the entities (ER) that are annotated or analysed in the text; the *components* as I have chosen to call them in this chapter. Genes and proteins usually have several different names which can incorporate lower- and uppercase letters, numbers and even symbols.

More advanced techniques are required when information extraction (IE), such as deriving relationship between entities (e.g. "protein A phosphorylates protein B" or "genes X and Y were not found to be co-expressed"), is attempted, as it is much more complex to extract semantic meaning from documents than to simply scan them for keywords as is most often the case in IR. An extension to relationship derivation is hypothesis generation, where de novo knowledge is sought from text. Swanson, an innovator in the field of hypothesis generation, constructed successful models as early as the mid eighties [254, 255]. Despite the fact that Swanson's models were assembled by himself and not automatically generated by a computer, his work shows how implicit connections can be made on the basis of seemingly unrelated literature. ARROWSMITH, an automated tool, partly based on Swanson's work has been publicly available since 1998 [238]. Swanson is still active in the field and recently published an article suggesting a relation between atrial fibrillation and overtraining-induced inflammation [256]. Excellent reviews by Jensen et al. [136] and Cohen & Hersh [54] describe biomedical literature mining and current approaches to it more thoroughly.

2.1.3 ENCODE

There has been a recent surge in efforts, not only to annotate the primary sequence data in databases, but to analyse variation in the sequence data, leading to even greater complexity in the information space of bioinformatics. Such efforts include the HapMap project [1] to map single nucleotide polymorphisms (SNPs) and the ENCODE project, which aims at functionally annotating 1% of the human genome to set the standard for methods and tools chosen to analyse the remaining 99% [58].

The data and analyses produced by such projects may not just enhance our current view of biology, but even paradigms such as the *"one gene/one enzyme"* theory[3] may be up for review as Carninci et al. conclude that most genes may have several gene products, many of which seem radically different and most likely have different functions [43].

Within the BioSapiens Network of Excellence, several European bioinformatics groups have compiled annotations for the ENCODE data and these have been integrated and published as described in the next section. The paper shows the diversity of annotations possible and how variation within loci needs to be taken into account when assigning function to genes and proteins. The methods used to integrate the data will be described later in section 2.3. The following paper serves as a showcase of the diversity of data types within proteomics and bioinformatics as well as that the huge amounts of sequence data produced are now being expanded with variation data on individual genes and proteins.

[3] Put forth by Beadle and Tatum and taught (in a refined version) in most textbooks on biochemistry.

2.2 Paper I

Proceedings of the National Academy of Sciences of the United States of America; 104: 5495-5500 (2007); doi:10.1073/pnas.0700800104

Implications of alternative splicing in the ENCODE protein complement

Michael L. Tress, Pier Luigi Martelli, Adam Frankish, Gabrielle Reeves, Jan Jaap Wesselink, Corin Yeats, Páll Ísólfur Ólason, Mario Albrecht, Hedi Hegyi, Alejandro Giorgetti, Domenico Raimondo, Julien Lagarde, Roman Laskowski, Gonzalo Lopez, Michael I. Sadowski, James Watson, Piero Fariselli, Ivan Rossi, Alinda Nagy, Wang Kai, Zenia Størling, Massimiliano Orsini, Yassen Assenov, Hagen Blankenburg, Carola Huthmacher, Fidel Ramirez, Andreas Schlicker, France Denoued, Phil Jones, Samuel Kerrien, Sandra E. Orchard, Ewan Birney, Søren Brunak, Rita Casadio, Roderic Guigo, Jennifer Harrow, Henning Hermjakob, David T. Jones, Thomas Lengauer, Christine A. Orengo, László Patthy, Janet M. Thornton, Anna Tramontano, Alfonso Valencia

The BioSapiens ENCODE consortium.

Abstract

Alternative pre-messenger RNA splicing is a mechanism that enables many genes to generate more than one gene product. Alternative splicing events that affect protein coding regions have the potential to create new protein functions and alternative splicing has been suggested as one explanation for the discrepancy between genome size and functional complexity. Here we carry out a full-scale study of the alternatively spliced protein isoforms annotated in the ENCODE pilot project. We find that alternative splicing is more frequent than commonly suggested in human genes. However, we also demonstrate that many of the potential alternative gene products will have markedly different structure and

function from their constitutively spliced counterparts. For the vast majority of these alternative protein isoforms little evidence exists for a functional role in the cell. It therefore seems unlikely that the spectrum of conventional enzymatic or structural functions is meaningfully extended through these alternative splice forms.

Introduction

Alternative RNA splicing, the generation of a diverse range of mature RNAs, has considerable potential to expand the cellular protein repertoire [173, 30, 240, 37, 191] and recent studies have estimated that 40-80% of multi-exon human genes can produce differently spliced mRNAs [291, 34]. The importance of alternative splicing in biological processes such as development [287, 292] has long been recognised and proteins coded by alternatively spliced transcripts have been implicated in a number of cellular processes, not all of which are advantageous [85, 263, 234, 280, 183].

The pilot project of the Encyclopaedia of DNA Elements (ENCODE) project [58], which aims to identify all the functional elements in the human genome, has undertaken a comprehensive analysis of 44 selected regions that make up 1% of the human genome. One valuable element of the project has been the detailing of a reference set of manually annotated splice variants by the GENCODE consortium [112].

While a full understanding of the functional implications of alternative splicing is still a long way off, this study is the first assessment of a systematically collected reference set of splice variants. We have been able to apply the best and latest computational methods and analysis tools and the most sophisticated information retrieval strategies to the GENCODE annotations. We are able to demonstrate that genes with alternative splice forms are in the majority in this set and that the alternative splice forms annotated here are likely to be an underestimation of the total pool of splice variants. A high proportion of the proteins coded for by these splice variants are predicted to be markedly different in structure and function from their constitutive counterparts.

Alternative splicing frequency

The annotation by the GENCODE consortium is an experimentally verified extension of the manually curated annotation by the Havana team at the Sanger Institute. The project has annotated 2,608 transcripts for 487 distinct loci and 1,097 transcripts from 434 loci are predicted to be protein coding. There are on average 2.53 protein coding variants per locus; 182 loci have only one variant while one locus, RP1-309K20.2 (CPNE1) has 17 coding variants (see Figure 1a).

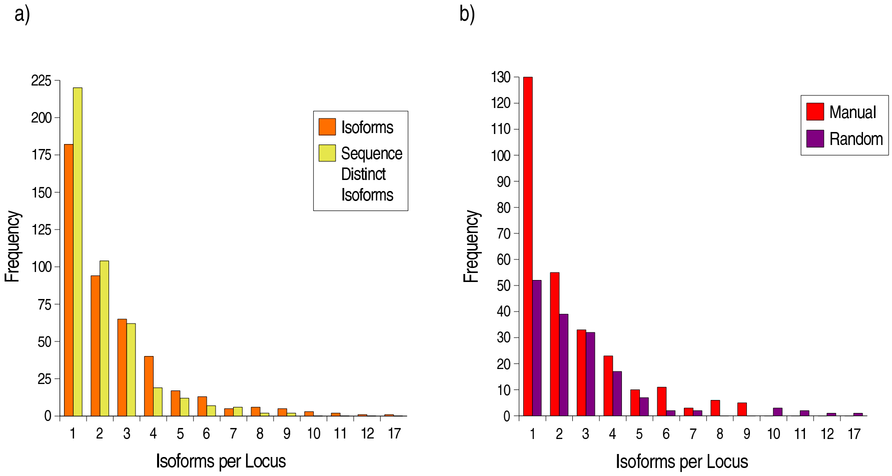


Figure 1. Isoforms per locus. Part (a) shows the number of isoforms per locus (orange) for the 434 loci in the set compared to the number of protein sequence distinct isoforms per locus (yellow). Part (b) shows the number of isoforms per locus in the manually selected regions (red) compared to the number of isoforms per locus in the regions selected by random stratified procedure (purple). The manual regions are rich in single isoform loci and this in part can be explained by the olfactory receptor cluster in manual pick number 9 (supporting online text).

A total of 57.8% of the loci are predicted to have alternatively spliced transcripts, although there are differences between those target regions chosen manually and those chosen according to the stratified random-sampling strategy [58]. The 0.5% of the human genome that was selected for biological interest has 276 loci and 52.1%of loci have multiple variants. The regions that were selected in the stratified random-sampling process have less loci (158), but more variants per locus (2.76) and 68.7% of the loci have multiple variants (see Figure 1b). This number is towards the higher end of previous estimates, but in line with recent reports [196].

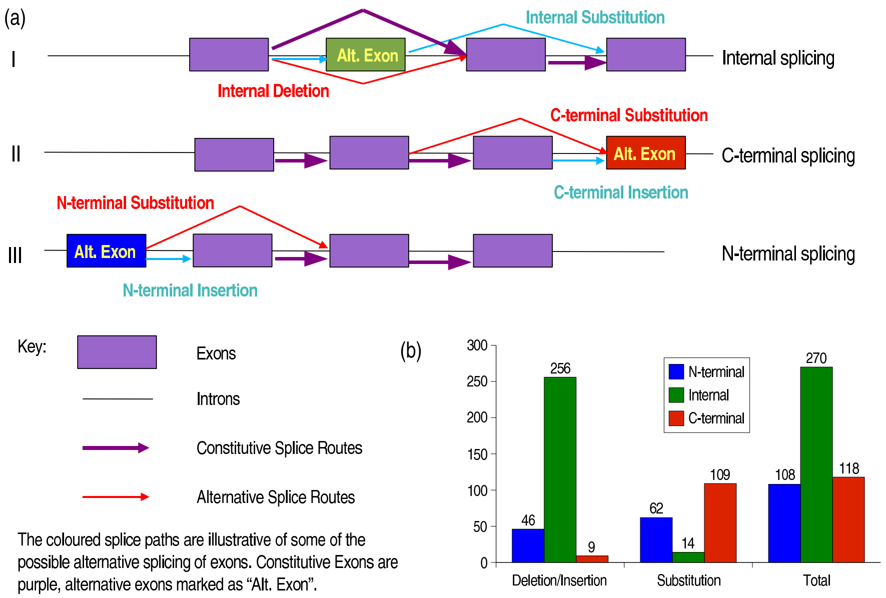
Much of the difference can be accounted for by the gene clusters in the manually selected regions for example in manual pick number 9 from chromosome 11 [257] there are 31 loci that are all labelled with the Gene Ontology [42] term "olfactory receptor activity". All 31 transcripts have a single isoform and alone account for half of the difference between the manual and random set. It is known that these 7 transmembrane helix olfactory receptors are found in large clusters, are recent in evolutionary origin and their loci rarely have variants. All the variants in this set have a single large coding exon and thus have more limited possibilities for alternative splicing.

A large proportion of the data set is composed of splice isoforms with identical protein sequences. These coding sequence-identical variants are alternatively spliced only in the 5' and 3' untranslated regions and form an interesting subgroup that may be under independent transcriptional control [298]. One locus, AF121781.16 (C21Orf13), has 11 alternative isoforms, all of which are sequence identical. This is not an isolated case: 230 of the 1,097 isoforms are identical, 25 loci have four or more identical isoforms and 15% of the loci with multiple variants code for nothing but protein sequence identical isoforms (Figure 1).

While there are a large number of identical sequences in the same locus, there are also identical protein sequences on separate loci, even though the selected regions only make up 1% of the human genome. Remarkably, protein sequence identical isoforms of the protein TEX28_HUMAN (Testis-specific protein TEX28) turn up four times in three different loci. There are two protein sequence identical isoforms in locus AC092402.6, another in a locus from the same clone (AC092402.4) and a fourth in a different clone (Z68193.2). All four instances came from same target region selection - manual pick number 6.

Gene mapping

Of the 434 genes in the data set, 417 loci have primary gene products that can be mapped to the Uniprot protein sequence database [288] (Table S1). Structures have also been resolved for a surprisingly high number of these genes; sequences from 42 different loci (almost 10%) have at least part of their structure deposited in the Protein Data Bank [26]. This reflects the effort that has been put into resolving the structure and function of human proteins. Locus GS1-273L24.4 stands out, it has two sequence-distinct splice variants (MTCP1 and MTCP2) and the entire structures of both proteins have been solved, in part because of the gene's role in T cell leukaemia [51].



The coloured splice paths are illustrative of some of the possible alternative splicing of exons. Constitutive Exons are purple, alternative exons marked as “Alt. Exon”.

Figure 2. Alternative splicing In part (a) we illustrate some of the potential types of alternative splicing. A splicing event can be internal as in row I, at the C-terminal, as in row II or at the N-terminal as in row III. In each case there are several paths splicing can take around an exon. The usual (constitutive splicing) path follows the black lines, while alternative splicing can miss out all or part of an exon, or can substitute on exon for another. Substitution of exons (leading to substitution of protein sequence) happens most frequently at the terminal ends of the mRNA sequence as comparisons of the splice isoforms in (b) shows. The most frequent splicing event is the removal or insertion of a whole exon within the mRNA sequence.

Functional and structural characterisation of alternatively spliced isoforms

In this part of the study we concentrated the analysis on the 214 loci that code for protein sequence distinct splice isoforms. Alternative splicing can take a number of forms (see Figure 2) and we classified the changes brought about by splicing events into 6 types. Since deletions and insertions cannot always be easily distinguished they were pooled. The results agreed with previous studies [194] - internal changes are almost always deletions or insertions of single or multiple exons and C-terminal changes tend to be substitutions of one or more constitutive exons for an alternative exon. Insertions or deletions are rare at the C-terminal (Figure 2). A number of substitutions in the protein sequence result from the translation of a different reading frame. For example in locus RP1-309I22.1 (TIMP3) alternative splicing between exon 4 and 5 leads to a frame shift in the fifth exon and means that the C-terminals of two variants are coded for from two different reading frames of the fifth exon. Another example is RP4-614O4.1 (ITGB4BP, Figure 3). The phenomenon of overlapping reading frames has been little studied in eukaryotes and there are only three functionally studied examples in humans. One example is INK4a/ARF21 [215] where different transcripts have CDS sharing 3' exons in different reading frames. Here we find that 23 separate loci code for variants with overlapping reading frames. The examples in this set may not be functional, but the findings suggest that the frequency of variants with overlapping reading frames might be somewhat higher than previously suggested [169].

Because of the nature of the manually selected regions, a relatively high proportion of loci code for proteins with trans-membrane helices (TMH). These TMH proteins are particularly interesting because their genes form clearly defined clusters, such as the 31 olfactory receptors on chromosome 11 [257], each with a single exon and no splice variants, and the natural killer cell immunoglobulin receptors clusters in manual pick number 1.

There are 41 loci that have isoforms with differing numbers of TMH. In most cases a single helix is lost relative to the principal sequence (the constitutive splice form), but there are also cases where 4, 5 and even 8 membrane sections are missing in the isoform. Interestingly, several genes appear to code for both soluble or transmembrane isoforms.

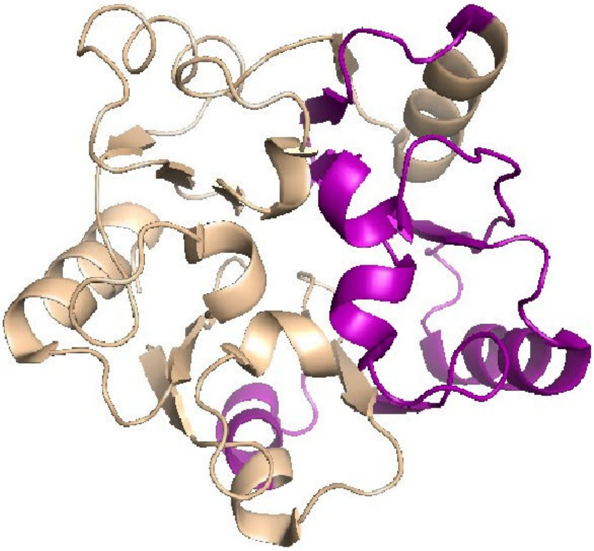


Figure 3. eIF6 structure. A structure exists with 75% sequence for the primary sequence of this locus (eIF6). 1g62A is a complex domain with pseudo five-fold symmetry and isoform 005 has a central substitution where 85 residues (marked in purple in the figure) are replaced by 60 non-homologous amino acids. The missing residues make up two of the five beta-alpha-beta propellers.

For example, locus AC006985.7 (UGT1A10) codes for two isoforms of UDP-glucuronosyltransferase 1A10. Isoform 002 has a short C-terminal substitution in place of the C-terminal 89 residue segment that contains a predicted TMH in the principal sequence, isoform 001. All 64 UDP-glucuronosyltransferases deposited in the SwissProt database [288] are annotated as monotopic membrane proteins and no natural soluble form is known. However, an engineered water-soluble form is reported [159]. If expressed, AC006985.7-002 would be the first soluble UDP-glucuronosyltransferase naturally encoded.

While splicing events often seem to splice out complete TMHs, there are cases where it is difficult to predict the resulting membrane topology. In locus AC129929.4 (TSPAN32), for example, the principal sequence (tetraspannin-32) has four

TMH, but the gene also codes for four different splice isoforms that each lose one membrane-spanning helix. In isoform 003 the N-terminal helix that acts as both a signal sequence and a membrane anchor [251] will be affected by an N-terminal substitution and in isoforms 005 and 012 the C-terminal TMH is lost, also through substitution. Isoform 014 lacks not just the N-terminal helix, but also the third TMH. This would leave the isoform as a protein with two membrane-spanning regions and with the first helix oriented in the opposite direction with respect to the other isoforms! All these cases evidently must force a change of structure or polarity, and all will result in a change of function if the protein is stable.

Of the 1097 transcripts, 219 were predicted to have signal peptides, accounting for 107 of the 434 loci. Unequivocal loss or gain of signal peptides can be seen in 12 loci. One obvious consequence of this is that localisation will not be conserved between isoforms. In eight loci the signal peptide loss/gain results from a substitution of exons at the N-terminus, as seen in RP1-248E1.1 (MOXD1) where one isoform loses 86 N-terminal residues including the signal peptide. This is coherent with earlier findings [194] that showed that most signal peptide gain/loss within alternative splicing products comes about through N-terminal exon substitution.

In AC010518.2 (LILRA3) signal peptide loss appears to be triggered by an N-terminal insertion. This results in the apparent internalisation of the signal peptide. Isoform 003 has a 17 residue N-terminal insertion ahead of the signal peptide that is predicted for the principal sequence (001). If the signal peptide is internalised in this isoform, the localisation of the protein will not be conserved. If this variant turns out to be expressed there is some evidence to suggest that its expression may be disease-associated the only supporting evidence for this isoform is in the form of ESTs from leukemia blood.

Definitions of protein functional domains can be extracted from a number of sources. We used the definitions from the Pfam database [20], both the hand-curated Pfam-A domains and the automatically generated Pfam-B domains. The start and end points of Pfam-B domains are less clearly defined, but including Pfam-B domains improves the coverage.

Splicing events occur within Pfam-A hand-curated functional domains in 46.5% of sequence-distinct isoforms and the figure rises to 71% if all Pfam defined domains are considered. Although this is a surprisingly high figure, it is still considerably less than might be anticipated. If the same number of splicing events occurred at random at the same exon boundaries, splicing events would be expected to occur inside Pfam-A domains in 59.8% of isoforms (84.8% in all Pfam-defined domains). It does seem that some form of selection has occurred, either in the generation or conservation of splice variants. As previously shown [158] this effect is not due to any correlation between domain and exon boundaries - we found no such correlation (supporting online text).

On occasion, splicing events leave out complete functional domains. In this set the effect was most marked with the immunoglobulin (ig) domain, a functional domain that is over-represented in the manually-chosen regions. Isoform 007 from locus AC011501.5 (KIR2DL4) is missing the N-terminal immunoglobulin domain and isoform 002 from locus AC010492 loses the second of four immunoglobulin folds that are present in the principle isoform 001. The isoforms in novel protein locus AC009955.5 also consist of strings of ig domains; one isoform (002) has a C-terminal truncation, whereas (003) has both an N and C terminal truncation.

The repeated use of splicing in altering immunoglobulin (Ig)-fold copy number of is particular interest when attempting to understand the involvement of ig-containing genes in developmental and immune system pathways. While the numbers of cases is undoubtedly influenced by the significant bias towards ig-like architecture in the manually-selected regions, it does suggest that this is not an isolated phenomenon and that it may occur in many other ig-fold containing proteins. It was also noticeable that no splicing event fell within an ig-domain in this set.

The infrequent variation in domain architectures that can be observed within the GENCODE annotations may be biologically meaningful and these loci are ideal candidates for further deeper study into the potential functionally relevant effects of alternative splicing.

Although these results do suggest that there is some favourable selection for splicing events that do not affect functional domains, there are still a large number of transcripts in this set where a splicing event does occur inside a domain and that apparently code for proteins with drastically altered structure and function. For example, in 49 of the 85 cases of alternative splicing where splicing events occur in regions covered by homologous PDB structures the resulting protein structure is likely to be substantially altered in relation to that of the principal sequence.

From gene expression to translation

Reverse transcription polymerase chain reaction (RT-PCR) experiments can confirm mRNA expression and it is indeed possible to find data for a marked number of loci. For example, both variants of AF030876.1 (MEC2P) have been shown to be expressed [157] and Tsyba et al. [260] confirmed the expression of a number of variants from locus AP000303.6 (ITSN1).

While it has been possible to confirm the expression of many alternative transcripts, it is important to know whether these genes are actually translated into proteins and whether the alternative splice isoforms with the most extreme deletions would become misfolded and quickly removed by the cell degradation machinery. Also, if the proteins are translated and fold properly, what functional role might they play in the cell?

It is clear from sequence comparison that a number of loci have alternative isoforms that must have radically different structures if they are to fold. Isoform 001 from locus RP11-247A12.5 (CRAT) has an 82 residue internal deletion in its acyltransferase domain (Figure 4) and isoform 005 from locus RP4-61404.1 (ITGB4BP) has an internal substitution that disrupts two blades of the stable alpha-beta propeller structure (Figure 3). Four separate loci from the serpin B cluster in random pick 122 have multiple alternative isoforms with large internal deletions or C-terminal substitutions (Figure 5). In all these cases it can be shown that the deletions would almost certainly mean that folding and function are severely affected.

In many of those cases where it was possible to build comparative models for the alternative isoforms we found that substantial rearrangement would be required to generate the models (Table S3). These isoforms could not be modelled by removing, adding or replacing a peripheral part of the protein structure: if these proteins are to fold properly and not aggregate some alternative structural and functional explanation must be invoked.

However, there is evidence that at least some alternative transcripts are expressed as proteins and can fold. Janssens et al. [133] showed that isoforms 008 and 011 of locus RP11-247A12.4 (PPP2R4) are translated in vitro and even though it is odd to imagine that proteins with radically altered folds are coded from the same genomic locus, this is exactly what happens in locus GS1-273L24.4 (MTCP1 and MTCP2). The two isoforms coded from non-overlapping regions of this gene have markedly distinct structures: MTCP-1 is a 117 residue filled beta-barrel, while MTCP-2 is a 68 residue 3-helix bundle (Figure S2).

Experimental evidence for functional differences between splice isoforms is harder to find. We were able to find just three concrete instances of functional differences between the splice isoforms in this set. Splice isoform 004 from locus AC034228.1 (ACSL6) has a sequence similar internal substitution corresponding to exon 11 in the transcript, a substitution that is one of the only two examples of mutually exclusive exon usage in the entire set. Kinetics assays show that this isoform has conspicuously different ATP binding affinities [271].

The three experimentally recorded splice variants of locus U52112.3 (IRAK1) coincide with the three coding sequence identical variants in the GENCODE set. IRAK1c (isoform 001 in locus U52112.3) has a 79 residue deletion in relation to the principle sequence (IRAK1, isoform 012) and it has been shown that IRAK1c differs from IRAK1 in that it does not undergo covalent modifications such as phosphorylation and ubiquitination upon lipopolysaccharide challenge and is not translocated to the nucleus [253]. IRAK1c is also the primary form in human brain tissue. The IRAK family proteins play critical roles in regulating innate immunity and the authors hypothesise that IRAK1c may keep brain tissue in a resting non-inflammatory state.

A further locus for which there is evidence of distinct functions is XX-FW83563B9.3 (TAZ). Vaz et al. [273] showed that alternative isoform 002, which has a 31 residue deletion from skipping the fifth exon, may actually be the correct principal sequence since it is the only isoform to have full cardiolipin metabolic activity.

This was not the only locus where doubts have been cast on the biological importance of the principal isoform. Recent work [224] has suggested that since cDNAs for many genes were cloned from tumour samples, the prevalent isoform may well have been coded from a tumour-specific splice variant rather than the mRNA sequence found in normal tissue. Indeed they found that tumour-associated splice forms were twice as likely to be represented in GenBank as the equivalent normal tissue-associated splice forms.

There has been abundant recent work associating alternative splicing with stresses incurred by cancer and other disorders [165, 152, 204], although rather than instigating the disease, in many cases the increase in expression of the aberrant variant may be a side effect of the general breakdown of cellular function. However, the importance of alternative splicing in cancer is such that cancer diagnosis can now be carried out using isoform-sensitive microarrays based on splice isoform profiles [38, 297].

It has been shown that at least two sets of alternative isoforms in this set are implicated in disease states, isoform 011 from locus AC051649.4 (TNNT3) in facioscapulohumeral muscular dystrophy [128] and isoform 006 of locus U52111.6 (L1CAM) in CRASH syndrome [95]. In addition, the mRNA supporting evidence for a number of variants was found exclusively in cancer cell lines (for example isoform 005 in locus Z97634.2, TMEM8, and isoform 003 from AC010518.2, LILRA3), suggesting that their expression may also be associated with disease states.

This analysis throws up several key questions, not least of which is whether the GENCODE-validated transcripts are comprehensive. While the set does include many known isoforms and uncovers numerous previously unrecognised variants, it seems very likely that many variants remain to be identified. For example, only four of the nine experimentally recorded isoforms [152] for locus XX-FW83563B9.3 (TAZ) are recognised, and the GENCODE set annotates just 3 of 6 Uniprot-recognised isoforms for locus AC011501.5 (KIR2DL4, Figure S6) and 1 of 4 Uniprot-recognised isoforms for locus AC129929.4 (TSPAN32). In fact, for the majority of loci that we looked at in detail, there were experimentally recognised variants that were not in GENCODE annotations.

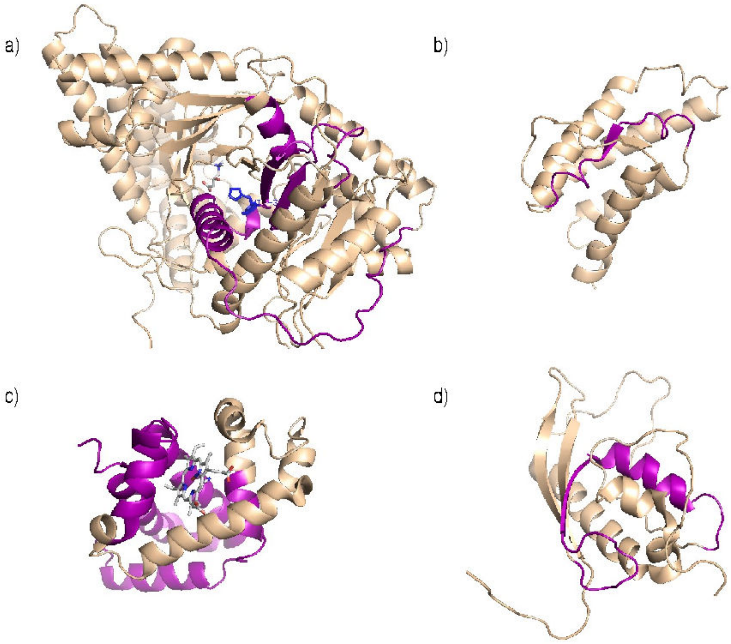


Figure 4. The potential effect of splicing on protein structure. Four splice isoforms from the data set mapped onto the nearest structural template. A. Carnitine O-acetyltransferase isoform 001 from locus RP11-247A12.5 mapped onto PDB structure 1s5oA. B. Interleukin 4 isoform 002 from locus AC004039.4 mapped onto PDB structure 2int. C. Hemaglobin delta subunit isoform 002 from locus AC104389.18 mapped onto PDB structure 1si4D. D. Sorting nexin 3 isoform 003 from locus RP3-429G5.4 mapped onto PDB structure 1ocuB. Structures are coloured in cream where the sequence of the splice isoform matches the structure, in purple where the sequence of the splice isoform is missing. The deletions will mean that the structures of these isoforms would require substantial reorganisation from the parent structure.

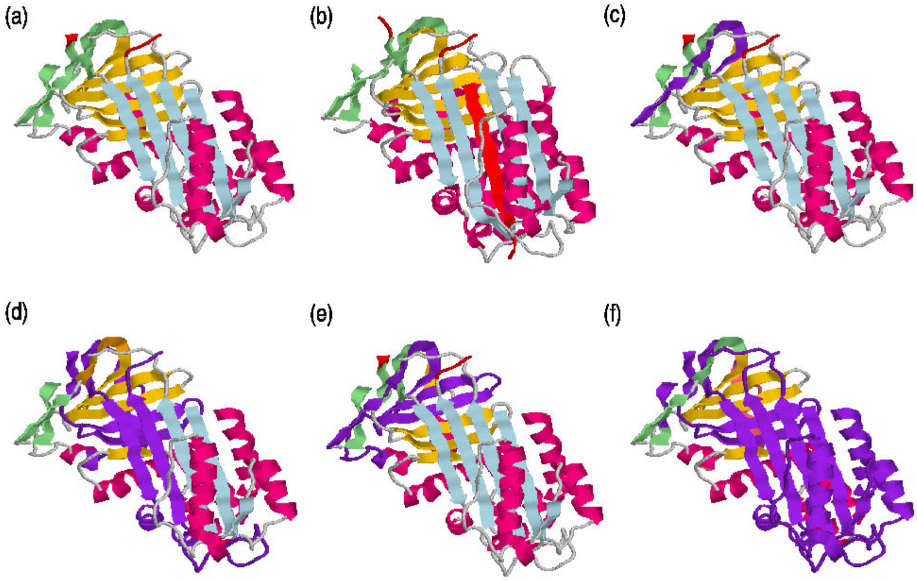


Figure 5. Clade B serpins. Serpins are primarily irreversible serine protease inhibitors. Their function is unique among protease inhibitors, they covalently inactivate their targets after undergoing an irreversible conformational change. Serpins have a complex fold containing a bundle of 8 or 9 alpha helices and a beta sandwich with three sheets (shown in silver, yellow and green in the diagram). They exist in an inactivated form that is regarded as being "stressed". Cleavage at the C-terminal end of the 20 residue RSL region, shown in red in part (a), allows a large conformation change in which the RSL region flips over and fits itself into one of the beta sheets. This form is shown in part (b) with the inserted RSL region in red. This is the form that is able to bind covalently and irreversibly inactivate the protease. Four serpin loci in the data set have multiple alternative isoforms with large internal deletions or C-terminal substitutions. In all eight of the isoforms it appears that the splicing is likely to cause the structure to fold in a moderately or substantially different fashion. In parts (c) to (f) we show four of the isoforms mapped onto the structure (1by7A). The sections deleted from the isoforms are shown in purple. Given that the complex structure of the inhibitor is vital to its unique function, it is puzzling that many variants exist that code for proteins apparently deleterious to this functionality.

Conclusions

We know the effect of splicing on the function for a few of the alternative isoforms in this set, but even in the cases described above we are still some way short of knowing their precise role in the cell. Here detailed and technically complex experimental approaches would be required and for most loci we can do little more than hypothesise as to the functional importance of splicing.

The study shows that alternative splicing is commonplace and the cross-section of alternative splicing events apparent at many different loci points to the potential versatility of alternative splicing in the creation of new functions. However, while alternative splicing has the potential to be an effective way of increasing the variety of protein functions, we see very little evidence of an increase in protein functional repertoire in this data set. In fact alternative splicing can lead to a wide range of outcomes, many of which may be undesirable. The dramatic changes in protein structure and function likely to result from the splicing of a large number of these variants suggest that many are likely to have functions that are potentially deleterious.

The standard path of protein evolution is usually conceived as stepwise single base pair mutations. Alternative splicing typically involves large insertions, deletions or substitutions of segments that may or may not correspond to functional domains, subcellular sorting signals or trans-membrane regions. The deletion and substitution of multiple exons seen in many of these transcripts suggests that splicing is not always a mechanism for delicate and subtle changes, but is a process that is rather more revolution than evolution.

What advantage is there to be gained in the cell from alternative splicing? The substantial changes evident in many of the alternative splice forms ought to disrupt their structure and function. Changes of this magnitude would normally not be tolerated because of the heavy selection pressure that must oppose such large changes [290].

While there may be as many evolutionary dead ends in alternative splicing as there are in standard evolutionary paths, one clear difference is that the organism seems to be able to tolerate these irregularities to some extent. It is possible that many of these splice variants lie more or less dormant within the gene and are only highly expressed as a result of some disease event. If splice variants in low numbers do not adversely affect the organism, the selection pressure against exon loss or substitution is reduced and the organism is able to tolerate the new variants. In this way large evolutionary changes can take place without significant repercussions, going some way towards explaining why so many of the alternative transcripts appear to encode proteins that are non-functional, at least in the classical sense.

2.3 Distributed systems

Integration implies segregation

As the short introduction to the diversity and abundance of biological data above shows, a single person or lab usually concentrates on only a few of the data types involved in sequence analysis. As we enter the post genomic era, the focus in systems biology is on integrating the diverse and dispersed data to form coherent pictures. The last sections of this chapter are dedicated to technologies and efforts to disseminate biological data.

Already when sequence analysis databases were in their infancy, people realised that at some point they would like to integrate the different data, and that this would be a tough task. In 1996, Robbins pointed out that the problems with database interoperability were threefold [222]:

Technical: Technical interoperability must be achieved, so that minimum functional connectivity can be assumed among participating information resources.

Semantic: Semantic interoperability must be developed, so that meaningful associations can be made between data objects in different databases.

Social: Social interoperability must occur, so that meaningful associations are made between data objects in different databases.

One could argue that the social component is in fact a integrated problem of both the technical and semantic problems, as there is no point in developing syntactic and semantic standards if no one uses them.

Since Robbins' report, XML and Internet technologies have solved the problem from the technical perspective, but the much harder problem of defining semantics for data integration remains largely unsolved. To a large extent, it is the social factor that is hindering advance, as a plethora of standards for data definition and sharing has been proposed (see 2.3.2). The problem lies in the lack of public acceptance of one standard above others, and now that so many data formats have seen the light of day, the real problem may in fact be how to map similar objects between data formats [248, 239].

The importance of being on-line

In 1980, the world of sequence analysis became a bit smaller when Dayhoff et al. published a sequence database, freely available to all over the telephone line [60]. Today, few scientists - especially within informatics - can imagine their jobs or even lives without an Internet connection. I have heard horror stories that the first biological sequences were published in paper format, and as I sit here and write this, I can see stacks of CDs reaching towards the ceiling in the neighboring office, CDs that contain sequence data, dating from times when download time and -cost for a GB of data were greater than good old fashioned snail mail. Today, access to the latest data and on-line methods is critical for success within the field of bioinformatics.

Federation vs. warehousing

Current sequence and annotation databases have a rapid turnover rate, and bioinformatics labs need to keep their data up to date to stay competitive. Database integration over the Internet can be achieved in many ways. *Distributed* or *federated* databases are those where the data sources are dispersed, but the data structure is unified to a degree. There are various degrees of interoperability for such databases ranging from total transparency, where the data sources are geographically dispersed, but the interface is the same, to interoperable databases where there is a similarity in data objects, but the interface is different. Integration of such data sources may be very easy, or very hard, depending on the level of coordination and standards their design implements. Making use of federated databases has the advantage that you are guaranteed to be using the most recent version of their data. The downside is that on-line sources occasionally go off-line, and having a single data source off-line may disrupt the projects and applications relying on them totally. An example application, relying on distributed databases for data is the CBS DAS browser described in section 2.4.

Many large projects, which rely on distributed source data, do not rely on federated access to those data, but assemble their own custom-made databases, integrating data from various sources. This approach is called *data warehousing*. As web-based (bioinformatics) resources are brittle in their nature, warehousing has been the mode of operation for many large scale integration projects destined for publication. An example of such an approach is the Inweb created at CBS for integration of protein protein interaction data (see section 4.3). The pros of having such a warehouse are that once implemented, having a local copy of all the data needed increases efficiency and ease of data manipulation. In other words: warehouses are self-contained. The cons are that a lot of work and logic is needed in the design phase to mold the various data sources into a coherent database. Such data repositories may also contain data sources that are not up to date and require more control over data versions.

As biological data exchange standards are maturing, some projects are combining federation and warehousing to create both a stable and up to date resource. Among those are the Ensembl and UCSC genome browsers [29, 118], which warehouse genome sequence data, along with annotations from various sources. Users can then add their own on-line data sources to be integrated to the browser view.

The greatest difference between these approaches is the robustness vs. concurrency. When you warehouse data - if you manage to get them into the warehouse - they remain there in a format of your choice and are easily and quickly available. Federation on the other hand requires that the data sources are available all the time as your integrated view of them will be compromised if and when they go offline. Federation gives you access to the latest data available at any time, whereas warehouses of data are usually compiled at intervals and may therefore contain obsolete data. Warehousing demands extra storage compared to federation. Federation, however, requires more calculations and network bandwidth at query time.

2.3.1 Identification, versioning and validation

Keeping your data up to date and knowing it is a complicated task. Sequence and annotation databases have a high turnover rate, constant updating and error checking is needed.

Accurate integration of bioinformatics data demands that we can verify that the data and annotations that we collect from distributed sources apply to the same sequences. This is not a trivial task; many projects and databases use different naming schemes and conventions. There is usually a handful of names for each sequence. To complicate matters even more, the same genes and proteins often have different names within the same database. The database entries get updated and many entry names have more than one version of a sequence tied

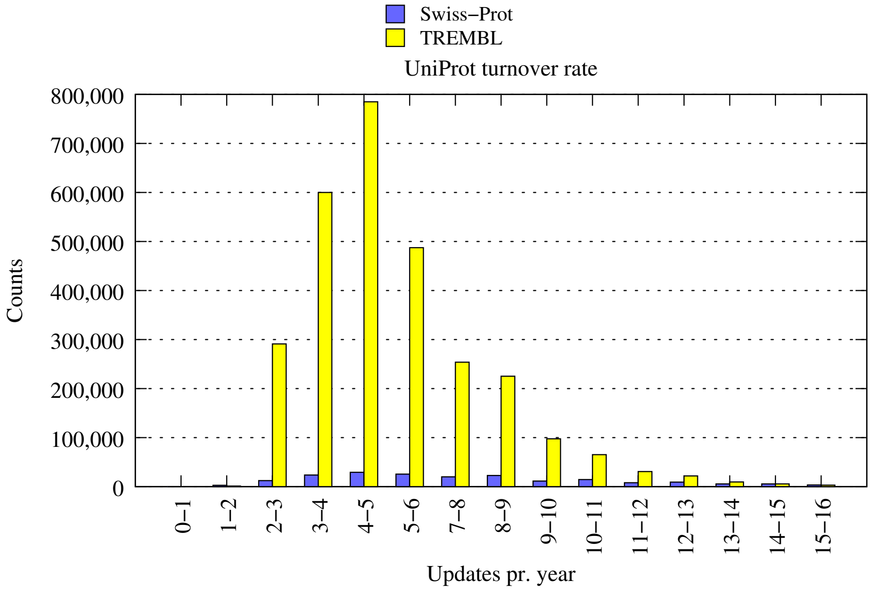


Figure 2.3: **The turnover rate of UniProt. Both the automatically annotated TREMBL and the manually curated Swiss-Prot undergo rapid changes. A protein entry in either database is likely to be modified on average 4-5 times every year. Annotation updates are much more common than sequence updates though, with only 1 sequence update in every 20 in Swiss-Prot and 1 in 5 updates being a sequence update in TREMBL.**

to them. As an example, the human insulin receptor, INSR_HUMAN (UniProt ID), has two sequences tied to it; the entry last modified in 2003. Figure 2.3 shows the average number of updates pr. entry in UniProt, with the average entry being updated about 4-5 times every year.

Additionally, the fact that we are constantly finding that the mapping between human genes and proteins is more diffuse that we thought: [43], section 2.2 and that there may be many more products per gene that previously thought.

As bioinformatics tasks become more and more automated, error checking, versioning and provenance grow more important.

MD5 checksums

For quite some time, people have used data digests, or *checksums*, to validate data being sent over a network. When a message is received or a download is completed, a checksum algorithm is put to work on the data and the results (usually a fixed length string) compared to the checksum downloaded separately for validation. This is a common approach to verify large downloads and that messages have not been tampered with. The MD5 checksum algorithm [221] takes as input an arbitrarily long message and outputs a 128 bit checksum.

In the case of biological sequences, ranging from a few residues to several thousand base pairs or amino acids in length, a fixed length checksum is a comfortable, computationally generated identifier. We have adopted this approach at CBS, indexing the major sequence databases with MD5 checksums. We store the external database identifiers in a synonym table to map from database identifiers to checksums and from there to the sequence. A very similar approach is described in [242]. The checksum approach has the advantage over other types of identifiers that it can be computed from the primary sequence and therefore no registry or sequence version information is needed; provided that the checksum algorithm used has a reasonably large and uniform output space, the checksums are unique and any two sequences will have the same checksum if and only if the sequences are really one and the same.

Life Science IDentifiers: LSIDs

The lack of standardization in naming convention within biology has sparked many projects, one of the most promising being the life science identifiers (LSIDs) project [53]. The aim of that project is to uniquely identify all entities, objects and services within life sciences with URIs[4]. LSIDs may well become the standard in biological entity identification as many projects and data repositories within bioinformatics already support LSIDs or have declared they will in the future: GenBank, PDB, Swiss-PROT, PubMed, DAS, BioMOBY, Gene Ontology, LocusLink, and Ensembl. As informatics moves toward Internet technologies and www standards, the use of URIs as identifiers will also become more and more accepted, within biology and other fields.

2.3.2 Exchange standards

On September 23rd 1999, a NASA spacecraft, the Mars Orbiter, fired its rockets for the last time before its planned landing on the surface of the red planet. To the mission control team's horror, the signal to the $125 million craft was lost and it presumably crashed into Mars. A subsequent investigation quickly found the cause of the failure; a simple error in the conversion from English standard units to metric ones.

While most of the time, errors in conversion between data formats and units do not result in multi-million dollar loss, more mundane examples of difficulties in data hacking see the light of day every day. During the writing of this thesis, I was asked to glue together two protein datasets, interaction data in one file and sequence similarity scores in another. I spent a whole day trying to understand why none of the data points were being linked as I tried to match the two. I finally realised that one of the datasets included a carriage-return (<CR>)

[4] "URIs or Uniform Resource Identifiers are classified as a locator or a name or both. A Uniform Resource Locator (URL) is a URI that, in addition to identifying a resource, provides means of acting upon or obtaining a representation of the resource by describing its primary access mechanism or network "location". For example, the URL http://www.wikipedia.org/ is a URI that identifies a resource (Wikipedia's home page) and implies that a representation of that resource (such as the home page's current HTML code, as encoded characters) is obtainable via HTTP from a network host named www.wikipedia.org. A Uniform Resource Name (URN) is a URI that identifies a resource by name in a particular namespace. A URN can be used to talk about a resource without implying its location or how to dereference it. For example, the URN urn:ISBN 0-395-36341-1 is a URI that, like an International Standard Book Number (ISBN), allows one to talk about a book, but doesn't suggest where and how to obtain an actual copy of it." From [284].

character [5] at the end of each data point and therefore my system did not match them as equal, even if they looked identical to the naked eye. This was not my first encounter with <CR> and probably not the last. Had the files been in a standardised format, this problem would have been avoided and my head would sport fewer gray hairs.

Due to the fact that standards need to be implemented on various levels, that is: *technical*, *semantic*, and *social* levels, they take time to be absorbed by the community. The best way to get a new standard out in the community is to keep it simple, allowing the community to start using it. The users will then shape the evolution of the standard, its success or failure. In this way, many practical solutions are compromises between ease of implementation and elaborate design; many ingenious data formats never catch on among the target users because of implementation problems. Given the dynamic nature of biological data (NCBI usually alters its on-line BLAST text output a few times a year), one might think that biological standards are a lost cause altogether. The trick seems to be to have standards as flexible as the data they describe, as oxymoronic as that may sound.

Microarray data represents complex analysis results, and as with all high-throughput experiments, the more data points you can integrate, the stronger your statistical evidence will be. In 2001, Brazma et al. proposed a data format for storage and exchange of microarray data [36] and now, five years later, this standard, Minimal Information About a Microarray Experiment (MIAME), is in widespread use, paving the way for research integrating microarray experiments from several labs using different hardware, experimental designs and interpretation methods [177]. Effective standards for biological data do exist.

As the web has evolved, we now have a plethora of standards to solve the technical issues of data integration in biology: DAS, Pedro, PSI-MI, mz-XML, AGML, SBML, SRS etcetera, etcetera. While many data formats are well designed and have shown potential, social factors have hindered their use in the general community. As examples, SRS [86] was a heavily used format for warehousing, implemented at EMBL among others. However, as the format and tools were commercial products, subject to licensing, the general public never

[5] "Carriage-return" is an end-of-line character, just like "line-feed" and "newline". Different computer systems interpret these end-of-line characters differently and while a Unix system, such as the one I work with most of the time, can see and understand a "carriage-return" character, it does not display it in any way, so a the strings "INSR_HUMAN" and "INSR_HUMAN<CR>" seem identical to a person working on a UNIX prompt, but the two are not equal, despite looking exactly the same, and therefore are not matched when compared.

embraced those tools. Likewise, the ASN.1 format[6] has proven useful in the human genome project, and is the primary storage and retrieval format in use at the National Center for Biotechnology Information (NCBI). ASN.1 failed to reach scientists in general, presumably because it uses binary encoding, which makes it hard to read message traffic without sophisticated tools.

PSI to the rescue

In 2002, the Human proteome organization (HUPO) founded the Proteomics Standards Initiative (PSI) to untangle the mess of data formats and exchange standards within proteomics [201]. The PSI has since then attacked the standardization problems in many subfields of proteomics and introduced standards and guidelines for data for microarray experiments, mass spectrometry, protein gel experiments and protein protein interactions.

Although the PSI initiative has proven very prolific and technically sound, it must still be considered immature, and mostly technical problems have been solved yet, while many problems with the semantics and social factors still remain. Having such a standard initiative is a big step in the direction of social acceptance of a standard though.

The PSI molecular interaction format (PSI-MI) [116] is now in widespread use. The format is very complex and allows for very detailed descriptions of molecular interactions. A huge problem in the initial use cases for the format was that the different databases providing data in the PSI-MI format, was that the different providers used format fields for different data types, there was no standard vocabulary of description, so a custom parser had to be designed for each data provider, just as if they all had different formats. The standard has now reached version 2.0, however, and these initial problems are being solved as the community usage matures and controlled vocabularies to describe fields evolve [202].

The plethora of standards emerging in bioinformatics reveal the scope of the problem of biological data integration. While it is obviously better to have well described standards for data sharing, we now face the problem of having too

[6] ASN.1 is a popular notation in communications, because of its richness as well as binary encoding of data which means lower bandwidth and transaction cost of messaging. ASN.1 was released in 1984, but as the standard has since been upgraded and modified repeatedly, the best source for information is the ASN.1 web page: `http://asn1.elibel.tm.fr/en/introduction/index.html`.

many such standards to choose from. Lincoln Stein, one of the designers of the DAS protocol, foresaw this in his 2002 article, “Creating a bioinformatics nation” [248], where he described the chaos of incompatible data being replaced or augmented by chaos from incompatible standards. I believe that choosing a “standard” standard from the soup is just as an important task for the PSI as defining new ones.

eXtensible Markup Language, XML

XML is a W3C[7] recommendation, and as a language a subset of Standard Generalised Markup Language (SGML), just like HyperText Markup Language (HTML). It is a language to describe data, but can also include data, and therefore, XML documents are self-describing and self-contained. It is an ideal language to describe structured, yet dynamic data and metadata.

The usefulness of XML has shown on the www, where there has been an XML explosion in the last decade. Virtually all information exchange is now handled with some form of XML formatting, and XML has become the common denominator of data sharing and processing on the web.

The impact of XML on Internet technologies, has been felt in the discipline of bioinformatics as most tools in use now offer some sort of XML input/output handling and virtually all new data formats and standard propositions are coded using XML syntax.

2.3.3 The Distributed Annotation System

As a partner of the BioSapiens Network of Excellence, 6th. framework EU project and the Danish Platform for Integrative Biology (DPIB), the Center for Biological Sequence Analysis has committed itself to disseminate its high-quality protein annotation methods in a standardised way. The exchange standard chosen was the popular Distributed Annotation System (DAS) [72]. Section 2.4 gives a thorough description of DAS and related infrastructure at CBS.

[7] The World Wide Web Consortium (W3C) is an international consortium where member organizations, a full-time staff and the public work together to develop standards for the World Wide Web. W3C’s stated mission is ”To lead the World Wide Web to its full potential by developing protocols and guidelines that ensure long-term growth for the Web.” From [277].

Having administered DAS sources and programmed DAS software for over 2 years now, I can safely state that, despite its simplicity, DAS is a very useful standard and will continue to be so as the user community keeps growing. The CBS DAS browser described in section 2.4 has now been appointed as an official BioSapiens DAS browser by the European Bioinformatics Institute (EBI) and the coordinators of the BioSapiens project. The BioSapiens version of the browser is available at `www.biosapiens.info`.

The ENCODE lateral work package to the BioSapiens project has also benefited greatly from the usefulness of DAS, as most groups within the network were able to annotate the ENCODE related sequences and distribute their annotations in a coordinated and coherent manner through DAS in a matter of days.

2.4 Paper II

Nucleic Acids Research; 33, W468-W470 (2005); doi:10.1093/nar/gki463

Integrating protein annotation resources through the Distributed Annotation System

Páll Ísólfur Ólason

Center for Biological Sequence Analysis BioCentrum-DTU, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark

Abstract

Using the Distributed Annotation System (DAS), we have created a protein annotation resource available at our web page: `http://www.cbs.dtu.dk`, as a part of the BioSapiens Network of Excellence EU FP6 project. The DAS protocol allows us to gather layers of annotation data for a given sequence and thereby gain an overview of the sequence's features. A user-friendly graphical client has also been developed (`http://www.cbs.dtu.dk/cgi-bin/das`), which demonstrates the possibilities of integration of DAS annotation data from multiple sources into a simple graphical view. The client displays protein feature annotations from the Center for Biological Sequence Analysis, CBS, as well as from the BioSapiens reference UniProt server (`http://www.ebi.ac.uk/das-srv/uniprot/das`) at the European Bioinformatics Institute (EBI). Other DAS data sources for protein annotation will be added as they become available.

Introduction

In recent years, numerous computational tools for gene and protein analysis have been constructed by various laboratories. Several such analysis tools have been created and published by CBS, many of which are available on-line for all users at the CBS web page: `http://www.cbs.dtu.dk`. The analysis results of such tools has led to an explosion in the amount of data in biological databases and available information there exists for biological sequences. Today, one of the major tasks of systems biology is to integrate as much of the experimental and computational information as possible and thereby gain biological insight into the properties and function of the macromolecules under observation. This means the assembly of several types of data, in various formats, dispersed around the face of the globe into a unified structure. This integration of on-line annotations is greatly simplified if the annotation services follow accepted standards. One such standard is the Distributed Annotation System (DAS) [72].

DAS services have existed for several years now. Version 1.0 of the DAS specification was released in 2001 and version 2 is under development. The DAS protocol is a simple http-based client-server system. A query on the form of a URL is made to the server, which replies with annotations for the sequence entry specified in the URL query. The server reply is XML formatted. The DAS web page (http://www.biodas.org) has both Perl- and Java-based server software for download. Client libraries in Perl and Java are also available.

The DAS specification was originally written with genomic sequences in mind, but the standard has proven itself flexible enough to handle protein data as well. Several annotation databases are now serving annotations using the DAS system, including Ensembl [124] , FlyBase [73], UniProt [15] and WormBase [48].

The flexibility and success of the DAS protocol has made it the annotation method of choice for the BioSapiens Network of Excellence, of which the CBS DAS server detailed here is a part. The various consortium members will in the near future deploy several DAS servers, which will serve protein annotations for the same UniProt sequences as the DAS server at CBS and all the data can therefore easily be integrated in a coherent manner.

The full list of query types that the DAS specification supports is beyond the scope of this document, we refer readers to the DAS web page and specification for detailed information and suffice to say that for queries on protein sequences, the most important queries are probably ”sequence” to which a reference DAS server responds with the full sequence and ”features” to which reference and annotation servers respond with feature annotations they store for a specified sequence identifier. An example query to the CBS DAS server is shown below.

Server infrastructure

At CBS, we have implemented a Perl-based DAS server, ProServer (`http://www.sanger.ac.uk/Software/analysis/proserver`), which accepts queries at the address: `http://genome.cbs.dtu.dk:9000/das`. We serve annotations for several of CBS's protein sequence annotation servers, which predict protein sorting (LipoP [143], NetNES [160], SignalP [24], SecretomeP [23], TargetP [80]), protein post-translational modification (NetAcet [149], NetPhos [31], NetOGlyc [142], NetNGlyc, ProP [74]) and protein structure and function (TMHMM [244]). Statistics and data source names for the individual methods are shown in table 2.1. The annotations served by the DAS server include: The start and end position of the feature annotated; the score from the prediction method that assigned the feature; a hyperlink to the web page of the prediction method with sequence information preloaded in the form input and possibly some further information.

In general, the annotations span all of UniProt [15], but are limited to phylogenetic subsets of the database, as the annotation methods are usually constructed with a specific phylogenetic group as a target (see the reference for each server for details). At the time of writing, the CBS DAS servers provide over 18 million protein annotations for over 1.5 million protein sequences from the UniProt database and we hope that this wide coverage makes our services of general interest to the scientific community. The predicted annotations include several highly cited methods e.g. SignalP and NetPhos, which are among the top 1% of the most cited papers in the scientific literature according to the Institute for Scientific Information, ISI.

The annotations are precalculated and the results stored in a relational database, allowing for fast retrieval and update of data.

Regarding the terminology of the predicted features, we have mostly used the nomenclature of the original prediction method. In some cases, we have modified the feature names to mimic the UniProt feature table, thus reflecting the reference database structure, allowing for easy comparison between the reference UniProt server and other annotation resources. It is quite conceivable that the vocabulary will be updated at a later point to make use of standard ontologies such as the Gene Ontology (GO) [10], so that post-translational modifications would be mapped onto GO "biological process" etc. The concept of the Sequence Ontology (SO) (`http://song.sourceforge.net/`) is highly relevant to this project, however the SO does not yet provide sufficient coverage of protein

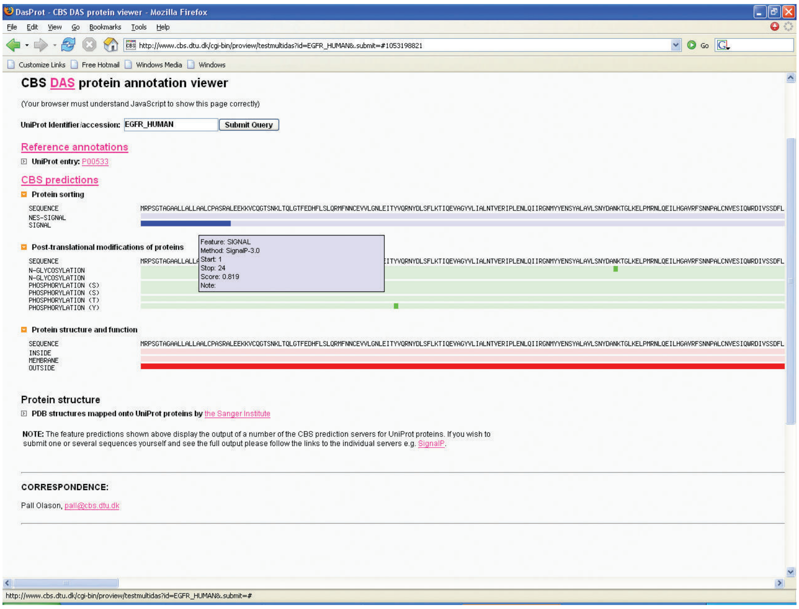
Method	Data source name	Organism coverage	Number of records	Reference
LipoP-1.0	lipop	G^{neg}	7,597	[143]
NetAcet-1.0	netacet	E	122,664	[160]
NetNES-1.1	netnes	E	1,945,054	[149]
NetNGlyc-1.0	netnglyc	H	137,800	
NetOGlyc-3.1	netoglyc	M	81,310	[142]
NetPhos-2.0	netphos	E	8,940,654	[31]
ProP-1.0	prop	E	127,553	[74]
SecretomeP-1.0	secretomep	E	58,318	[23]
SignalP-3.0	signalp	E, G^{pos}, G^{neg}	1,189,706	[24]
TargetP-1.01	targetp	E	750,111	[80]
TMHMM-2.0	tmhmm	A	5,086,476	[244]
All above combined	cbs_total		18,447,243	

Table 2.1: **Annotation methods provided by the CBS DAS system. The annotation methods are specific to the following phylogenetic groups: ”A” stands for all proteins, ”E” for eukaryotes, ”G^{pos}” for gram positive bacteria, ”G^{neg}” for gram negative bacteria, ”H” for human and ”M” for mammals. The data source name is the name of the particular annotation method on the DAS server.**

sequence attributes, such as post-translational modification, to be useful for our purposes.

A query example

When querying a DAS server for annotation, one must append the data source name (DSN), along with a query type and a sequence identifier to the address of the server. For example: If we wish to ask for annotations from the SignalP signal peptide prediction method [24] for the protein EGFR_HUMAN we first append the DSN for that method (“signalp”, see table 2.1). Then we use the ”features” query to ask for feature annotations and identify the sequence as a ”segment”. The whole query string thus looks like this: `http://genome.cbs.dtu.dk:9000/das/signalp/features?segment=EGFR_HUMAN`.


DasProt - CBS DAS protein viewer - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://www.cbs.dtu.dk/cgi-bin/proview/testmultidas?id=EGFR_HUMAN&.submit=#1053198821
Go
Customize Links
Free Hotmail
Windows Media
Windows

CBS DAS protein annotation viewer

(Your browser must understand JavaScript to show this page correctly)

UniProt Identifier/accession: EGFR_HUMAN **Submit Query**

Reference annotations

⊞ **UniProt entry:** P00533

CBS predictions

Protein sorting

```
SEQUENCE       MRPSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEVVLGNLEITYVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNMYYENSYALAVLSNYDANKTGLKELPMRNLQEILHGAVRFSNNPALCNVESIQWRDIVSSDFL
NES-SIGNAL
SIGNAL
```

Post-translational modifications of proteins

```
SEQUENCE       MRPSGTAGAALLALLA                    EITYVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNMYYENSYALAVLSNYDANKTGLKELPMRNLQEILHGAVRFSNNPALCNVESIQWRDIVSSDFL
N-GLYCOSYLATION
N-GLYCOSYLATION
PHOSPHORYLATION (S)
PHOSPHORYLATION (S)
PHOSPHORYLATION (T)
PHOSPHORYLATION (Y)
```

Feature: SIGNAL
Method: SignalP-3.0
Start: 1
Stop: 24
Score: 0.819
Note:

Protein structure and function

```
SEQUENCE       MRPSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEVVLGNLEITYVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNMYYENSYALAVLSNYDANKTGLKELPMRNLQEILHGAVRFSNNPALCNVESIQWRDIVSSDFL
INSIDE
MEMBRANE
OUTSIDE
```

Protein structure

⊞ **PDB structures mapped onto UniProt proteins by** the Sanger Institute

NOTE: The feature predictions shown above display the output of a number of the CBS prediction servers for UniProt proteins. If you wish to submit one or several sequences yourself and see the full output please follow the links to the individual servers e.g. SignalP.

CORRESPONDENCE:

Pall Olason, pall@cbs.dtu.dk

http://www.cbs.dtu.dk/cgi-bin/proview/testmultidas?id=EGFR_HUMAN&.submit=#

Figure 2.4: **The CBS protein DAS viewer. The browser interface is very simple, it has only one form field and the graphical tracks show the annotations for a given UniProt protein. Additional information for individual features is shown in a pop up help window when the mouse is pointed at the feature.**

CBS DAS viewer

As the raw XML output of DAS servers is not very suitable for browsing of feature annotations, we have developed a client viewer to allow visualization of CBS DAS annotations in a simple graphical way. This viewer is publicly available at `http://www.cbs.dtu.dk/cgi-bin/das`. All the user is required to do is to input a UniProt accession number or identifier. The viewer then collects the annotations served by the CBS DAS servers, along with annotations from a UniProt reference DAS server at the EBI (`http://www.ebi.ac.uk/das-srv/uniprot/das`) for that particular sequence. All the annotations are then displayed as aligned graphical tracks, allowing for easy inspection of features along the length of the protein. Additional information about the annotations is shown in a pop-up window when the user points the mouse to an annotation track.

This is the first time CBS provides a composite graphical display of several of its protein prediction methods simultaneously, which the users of CBS prediction services may find interesting. Some types of feature annotations carry a hyperlink in the XML payload. When the user clicks on a graphical track for such an annotation, the CBS DAS protein viewer will open a new browser window, following the hyperlink. The graphical tracks can also be folded and expanded to allow simplified overview. A screenshot of the client in action can be seen in figure 2.4. The client demonstrates how easily different data sources can be integrated using the DAS system. We plan to incorporate relevant DAS protein annotation resources into the graphical client as they appear. At the time of writing, only one external DAS source is incorporated in the view; a resource where RCSB Protein Data Bank [68] structures are aligned upon UniProt entries, provided by the Sanger Institute (`http://www.sanger.ac.uk`).

Acknowledgements

The author wishes to thank Søren Brunak, Kristoffer Rapacki and Hans-Henrik Stærfeldt at CBS, as well as the anonymous reviewers for helpful comments and suggestions. This work was supported by a grant from the BioSapiens Network of Excellence (NoE), FP6, contract no.: LSHG-CT-2003-503265 and the Danish Platform for Integrative Biology with Focus on Systemic Proteomics - DPIB, funded by the Danish National Research Foundation.

2.5 Web services and workflows on the net

2.5.1 The semantic web

In an article in *Scientific American* from 2001, the inventor of the world wide web, Tim Berners-Lee[8] , produces an imaginary scene where two people schedule a series of meetings, book a venue and rearrange less important coinciding tasks automatically with the use of web agents[9] .

It may be hard for people to imagine that the chaotic mess that is the Internet will acquire such logic and functionality in the near future. But who, only ten years ago, could have predicted the impact that mobile phones, handheld wireless devices, and last - but not least, the world wide web would have on our everyday lives.

Berners-Lee's vision of a semantic web really isn't far-fetched; what seems to be artificial intelligence in agent software, making arrangements for users on-line, is really simplified with syntax standards and formal descriptions of services that are already available on the web. The whole point of the semantic web is to automate tasks; to use registries to discover distributed services, ontologies to determine the nature of these services and the entities and objects used.

Using protocols such as web services description language (WSDL), simple object access protocol (SOAP) and resource description framework (RDF)[10] , along with dozens of other web object description standards, on-line analysis services can be described in a machine-understandable fashion and then there is little logic needed to connect the dots and create smart workflows such as ordering lowest-possible airline fares, making an appointment at your dentist, and writing up a list of groceries that your Internet-connected refrigerator is running low on. All the case scenarios described here above are already existing technologies, not sci-fi fantasies. The technology is available. The greatest problems are, however, to find the relevant services and to decide whether to trust them or not. The trust issue is related to Internet security, and I will not discuss that field in

[8] Berners-Lee invented the web as a means for scientists to easily share papers and documents on-line, while working at the European Organization for Nuclear Research (CERN).

[9] Agent: A piece of software that runs without direct human control or constant supervision to accomplish goals provided by a user. Agents typically collect, filter and process information found on the web, sometimes with the help of other agents. From [27].

[10] RDF and SOAP are W3C recommendations, WSDL is expected to become a W3C recommendation as it matures.

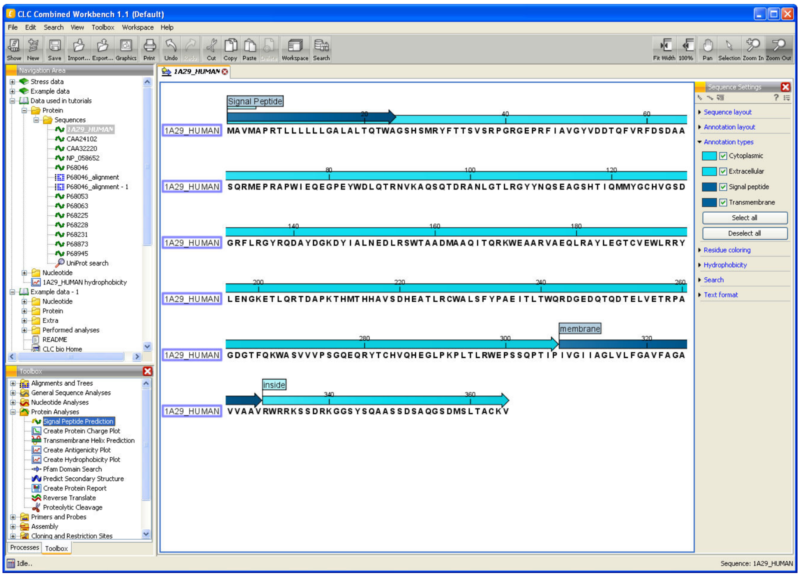
here. The problem of finding relevant services for the workflow or arrangements you wish to perform on-line has been discussed for some time in the literature, and while there are protocols for finding relevant web services (e.g. universal description, discovery and integration, (UDDI), protocol [265]) and some have suggested to simply crawl all of the web for them [94], using "knowbots". At the moment, the practical solution is to use a registry of available services along with a description of their methods, input and output objects as suggested by Lincoln Stein in [248]. Indeed, my own experience from the DAS work in the BioSapiens network is that a registry is an easy-to-use, practical, and, if standardised, easily automated way of keeping track of relevant data and analysis methods on the web.

The requirement of being able to access analysis services on-line is felt at CBS, which is a part of several projects, that require on-the-fly computations and access to databases. In anticipation of these requirements, we have set up web services for most of the protein annotation servers available at the CBS analysis site (`http://www.cbs.dtu.dk/services`). While these services are strictly a pilot test, they were successfully set up and are currently being used by some of our partners. Figure 2.5 shows how client software can access CBS' web services and run distributed workflows with sequence analysis performed in real time.

Ontologies

In an attempt to glue data and standards together, several efforts have been initiated to decide on a controlled vocabulary to describe the objects and entities allowed in biological data formats. Gene Ontology (GO) [111] is the best-known of these and the model for most new biomedical ontology projects. While GO is in many ways useful in describing data, its dynamic nature and constant development make it a difficult platform to use for automated tasks as it is of course hard to use a standard, if the standard is in constant flux.

Many new projects have followed the GO and as with syntax standards, now it seems as if the future challenge will not be to use a controlled vocabulary, but to select a single one and implement as *the* controlled vocabulary. Open biomedical ontologies (OBO) [197] is a consortium of ontology projects, ranging from genomics to phylogenetics to phenotypes. The OBO aims at standardisation of its member vocabularies and, as such, has the potential to become an ontology of ontologies; a registry where researchers working with data integration can fetch controlled vocabularies, not just for a data set from one domain,



```
1A29_HUMAN  MAVMAPRTLLLLLLGALALTQTWAGSHSMRYFTTSVSRPGRGEPRFIAVGYVDDTQFVRFDSDAA
1A29_HUMAN  SQRMEPRAPWIEQEGPEYWDLQTRNVKAQSQTDRANLGTLRGYYNQSEAGSHTIQMMYGCHVGSD
1A29_HUMAN  GRFLRGYRQDAYDGKDYIALNEDLRSWTAADMAAQITQRKWEAARVAEQLRAYLEGTCVEWLRRY
1A29_HUMAN  LENGKETLQRTDAPKTHMTHHAVSDHEATLRCWALSFYPAEITLTWQRDGEDQTQDTELVETRPA
1A29_HUMAN  GDGTFQKWASVVVPSGQEQRYTCHVQHEGLPKPLTLRWEPSSQPTIPIVGIIAGLVLFGAVFAGA
1A29_HUMAN  VVAAVRWRRKSSDRKGGSYSQAASSDSAQGSDMSLTACKV
```

🟦 ☑ Cytoplasmic
🟦 ☑ Extracellular
🟦 ☑ Signal peptide
🟦 ☑ Transmembrane

Figure 2.5: **CLC Bio's workbench, a tool to integrate sequence annotations from distributed web services. Here shown with two of CBS' web services loaded, SignalP [24] and TMHMM [244], performing computational sequence analysis on the fly.**

but for all domains of their research data. The ontology lookup service [59] is an interface to search ontologies that conform to the OBO standard.

Automated tasks

Today, workflow design in bioinformatics has reached so much maturity that one can call a registry, find an appropriate service, based on input and output, and assemble a workflow for your data. Several projects have been initiated to aid the construction of such pipelined workflows in biology: BioMOBY [285] and myGrid [250] focus on the networking and workflow enactment, while graphical clients such as Taverna [125], BlueJay [261], biowep (`http://bioinformatics.istge.it/biowep/index.html`), REMORA [44], and others allow users to create custom workflows and run them.

The building blocks are there and once the user interfaces for these services become more robust and intuitive this type of automation is sure to become widespread.

2.6 Where to?

Biological sequence data is being produced at ever-increasing rates and there is constant need to counter the growth with faster computers and more efficient algorithms for analysis. The problem of data diversity is more difficult to solve. In this chapter I have (I hope) produced a picture of the growing amounts and diversity of biological data. The trend in bioinformatics is towards a systems view of the cell; a bird's eye view that demands integration of all possible (and impossible) data types and even untyped data such as free text, that pertain to the components of these systems. As data integration implies that we have disparate data sources, this chapter has also described data distribution systems for biological information and my work in that field.

The informatics community has realised that something needs to be done to make sense of all the information and there are countless efforts towards data formats, information sharing standards and even an improved world wide web. The future challenge will be to pick one standard above others, or even adding a new layer of abstraction and start defining standards for standards.

As with other informatics disciplines, bioinformatics is riding the wave of Internet technologies, providing on-line access to methods and data. The world wide web is becoming more than just a storage of documents. It is becoming an on-line lab, enabling scientists to run analyses directly on their screen. The use of the web as a tool, rather than just static data, requires that we think of information on the web in a new way.

Chapter 3

Interconnections

Most proteins do not perform their function alone, but in concert with other proteins, whether as part of the same obligate complex or during a transient chemical reaction. Knowledge of protein-protein interactions will undoubtedly answer many of our questions regarding the function and properties of the interacting proteins. Several methods have been devised in the past few years to predict protein-protein interactions, with a varying degree of success. The high false positive rate of available experimental data, the lack of decidedly non-interacting protein pairs for use as negative data and lack of understanding of the true driving force behind protein-protein interactions, make the body of work published so far hard to assay and benchmark. In this work, we point out possible pitfalls in the handling of protein-protein interaction data with the aim of creating a prediction method.

Having introduced the components of genes and proteins and the annotations tagged onto them by distributed analysis services around the globe, it is now time to start connecting the dots. To gain a bird's eye view of the systems that the components are a part of; to analyse the whole instead of the individual, we need to start making links. The links can be based on the annotations introduced in chapter 2 or some other type of evidence.

Our ultimate goal is a systems' view of the cell's proteome. The natural backbone of such a system is the interactome, which describes physical interconnections between the protein components. This chapter describes efforts to produce such a scaffold. We then proceed to attach annotations onto the interactome and analyse it as a whole in chapter 4.

3.1 What are protein-protein interactions?

The most direct cellular relationship possible between any two sequences is direct physical contact. In proteomics, this means protein-protein interactions (PPIs). There are however several types of PPIs possible; there are transient types such as enzyme-substrate reactions, obligate complex formations and in between those types we have dynamic interactions, such as cyclin binding, demonstrated to be under dynamic control of the cell cycle [61].

3.1.1 Physical interactions

Physical evidence of protein-protein interactions can be derived in many ways, with two experiment types being the greatest contributors: yeast two-hybrid (Y2H) screens and complex purification followed by mass spectrometry (CP). Protein microarrays are a promising, but immature, addition to the field.

Yeast two-hybrid assays

This experiment type was conceived by Fields and Song [89] and relies upon the modular nature of transcription factors, i.e. that they are composed of a DNA binding domain and an activation domain. The two proteins being analysed for interaction are hybridised, one to the DNA binding domain and the other to the activation domain. If the two proteins, usually dubbed *bait* and *prey*, interact, the transcription factor becomes functional and a reporter gene gets transcribed. The yeast two-hybrid (Y2H) procedure is thus performed in vivo and is the first of two technologies that have been performed to map the physical interactions of an organism on a large scale [127, 266].

Complex purifications

The second experiment type used to derive physical interactions between proteins on a large scale is complex purification, followed by mass spectrometry (CP) [205]. In this approach, bait proteins are tagged with a molecule or chemical, over-expressed in cells, and then protein and its binding partners are purified. After separation with gel electrophoresis, the proteins are identified with mass spectrometry. A common way to isolate the bait is to tag the it with an epitope and use an antibody to precipitate it and its stable binding partners. This approach can reveal all the constituents of a complex in one experiment, but it is hard to say which of the complex proteins actually interact physically (see section 3.1.3 below).

Protein chips

Protein chips, or protein microarrays, [174] promise to enable high-throughput analysis of protein binding. The concept is similar to DNA/RNA microarrays, where the sequence is spatially fixed on a plate surface and binding partners are applied in solution. Protein array technology, however, is still immature as proteins are much more variable in chemistry and 3-dimensional structure than nucleotide sequences, making it hard to reproduce the true character of proteins as a spot on an array [156].

3.1.2 Implicit interactions

Biological experiments often imply interactions between genes and proteins despite there is no direct physical evidence for those reactions. Several types of such knowledge-based evidence have been put forth lately. The next section will touch upon the most commonly mentioned types of indirect interaction evidence.

Genetic evidence

With the ongoing analysis of the vast amount of genomic information available, several types of empirical data have been shown to be correlated to PPIs. Qin et al. [214] showed that the evolutionary classification of proteins is correlated with their tendency to interact, implying an evolutionary synergy among interacting proteins. In a similar vein, several studies have revealed that interacting protein pairs co-evolve and that the alignment distance matrices calculated for the proteins can be used as an indicator of their tendency to interact [102, 207]. Further, building on the phylogenetic trees of interacting proteins, correlated mutations have been used to predict the contact residues of interacting proteins [208]. Two studies have revealed that gene fusion events in one organism indicate that the two constituent proteins interact in other organisms [176, 81]. Direct, physical evidence for PPIs has also been mapped between organisms using sequence homology between the interacting protein pair and a similar pair in a different organism. Such interacting homologs are dubbed *interologs* [88, 146].

The STRING database [186] integrates many such knowledge-based associations into a unique, benchmarked score for each proposed interaction, as does the approach by Lee et al. [166]. Valencia and Pazos have written an excellent review of such knowledge-based methods: [267].

Co-expression

Co-expressed proteins also show a higher tendency to interact than proteins that are not co-expressed. This is true especially in the case of permanent complexes such as the ribosome and proteasome [130, 131]. Studying the dynamics of complex formation, de Lichtenberg et al. postulate that in many cases, the expression of single proteins controls the assembly of a whole complex [61].

Pathway co-occurrence

Proteins that take part in the same pathway in the cell are often annotated as interacting. Thus von Mering et al. use co-occurrence of proteins in KEGG pathways as benchmarks and validation of protein associations in the STRING database [186].

3.1.3 Public data

The abundance and usage of protein-protein interaction (PPI) data has exploded in the field of bioinformatics recently. The two largest sources of PPI data, the Y2H screens and CP experiments, have produced thousands of links in proteomics research.

As a model organism, the budding yeast *S. cerevisiae* is, by far, the best covered organism, both with regard to proteome coverage as well as absolute numbers of published interactions [266, 127, 97].

While PPI data holds much promise for functional genomics and systems biology, this data source has shown itself very biased and error-prone [276, 14]. Therefore, there is as much, or more, work published that focuses on cleaning and benchmarking such data, as there is work that makes use of them in integrated research.

Reliability of PPI data

As stated above, there are mainly two sources of experimental PPI data; the yeast two hybrid assay [89] and complex purification and identification by mass spectrometry [205]. A large concern is the lack of overlap between Y2H and

CP data. Several studies have shown that the overlap in the accumulated interaction data in yeast is less than 5% [14, 276]. This discrepancy may indicate that the two experiment types capture different types of interactions, where CP mainly produces stable complexes and Y2H yields more transient, binary interactions [134, 98, 5, 147]. Indeed, Edwards et al. found that Y2H assays did not report any of interactions between proteins in the proteasome [77], and conversely, several kinase-substrate reactions were identified in high-throughput Y2H studies [266, 127], but not found in the large-scale complex purification studies of Gavin et al. [98] and Ho et al. [119].

Spokes and matrices

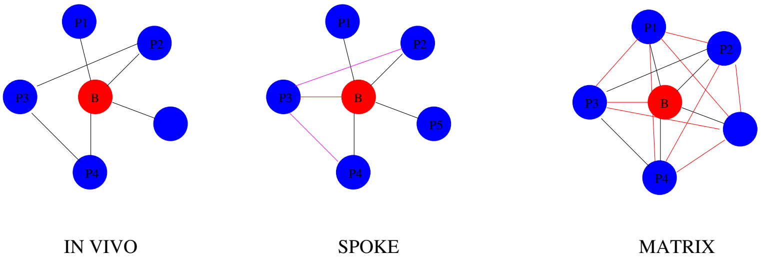


Figure 3.1: **Spoke and matrix models of CP experiments. This diagram depicts a fictional complex, where proteins are physically connected as the leftmost figure shows. In the spoke model (middle) interactions are derived between the bait (red) and all prey proteins (blue). In the matrix model, interactions are inferred between all proteins, regardless of them being baits or preys. For the complex in the diagram the spoke model has 4 correct PPIs, one false positive PPI and is missing two PPIs. The matrix model captures all the PPIs, but at the cost of adding nine false positive PPIs.**

When analysing CP data for complex pulldowns, it is hard to figure out which of the proteins, that are co-purified, actually interact physically. Two models are used to infer the interactions from a complex [12]: the *spoke* and *matrix* models.

The matrix model: A physical interaction is assumed between all pairs of co-purified proteins.

The spoke model: A physical interaction is inferred between the bait protein and all the prey proteins, but not between pairs of prey proteins.

Neither PPI derivation is perfect, but rather a compromise and one must choose accordingly. Do you want to exclude as many false positives as possible or do you want to maximise data yield, at the cost of false positives?

The numbers of derived interactions from spoke and matrix data are $n - 1$ and $\frac{n^2-n}{2}$, respectively, where n is the total number of proteins purified.

It has been argued that the spoke model depicts the pairwise physical interactions in a complex three times more accurately [12], but generally speaking it is hard to assign correct pairwise interactions based on CP data, as the prey proteins may just as well be attached to one another as to the bait protein. If one hypothetically uses all proteins found in a complex as baits, one at a time, it seems logical, that it will generally be the same proteins that are pulled down each time and therefore the matrix model can be viewed as an extrapolation of the spoke model for hypothetical assays of complex pulldowns over all proteins found in a complex. In addition, the spoke model only has $2/n$ times the interactions of the matrix model, where n is the total number of proteins in the experiment, bait and preys, so the data loss from matrix to spoke is considerable.

There is a flip side to that data loss, as the amount of pairwise PPIs derived by the matrix interpretation grows as the square of the number of proteins purified. This results in a combinatorial explosion for large complexes. As an example, the MIPS database identifies almost 140 proteins as taking part in ribosome formation. If a pairwise interaction is assumed between all those proteins, the resulting number of derived ribosomal PPIs is $\frac{140 \times 139}{2} = 9730$. In this way the number of ribosomal interactions would amount to a large percentage of the total interactome. I will return to this discussion when discussing data preparation and potential pitfalls of PPI prediction later in this chapter.

Databases

A number of databases designed to store and serve PPI data have been published. The European collaboration database, Intact [117], has grown to be the largest PPI database publicly available. The Intact effort warehouses data from the other databases, and usually there is considerable overlap in the data stored in the databases. Two highly cited efforts are DIP [230] and BIND [100], which have stagnated a bit in the last couple of years. A novel resource is the HPRD

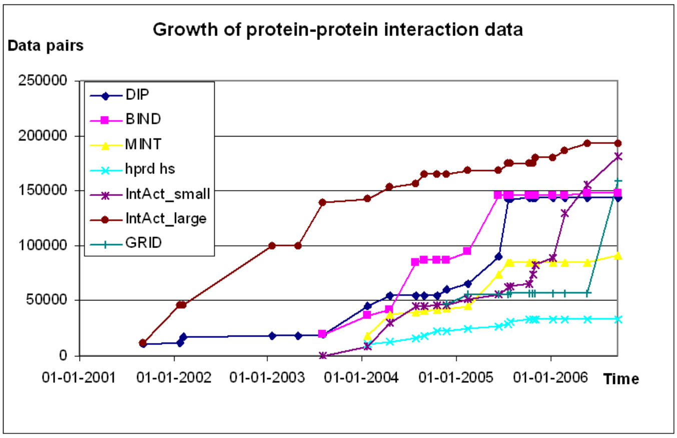


Figure 3.2: **Major PPI databases and their growth rate measured in the number of PPI pairs contained. (Statistics and figure by Olga Rigina, olga@cbs.dtu.dk.)**

database [190], which is by far the largest source of human PPI data, as it is derived by manual curation efforts. All these huge PPI resources and many more have been warehoused at CBS for some years now, allowing for the integration and analysis of a huge body of PPI data, paving the way for direct PPI research as described in this chapter and for indirect usage of PPI data in projects such as disease gene candidate prioritisation (see chapter 4).

Journal text

As described in the previous chapter, the volume of published articles is rocketing and researchers have a hard time to follow the body of current research, even in a narrow field. Several groups have initiated text mining efforts to extract functional relationships between protein pairs. While such relationship data is noisy and often the nature of the interactions fuzzy, it is clear that research in this greatest resource of scientific data is important. See [186, 120, 295] to name a few approaches.

3.2 Prediction of protein-protein interactions

While protein-protein interconnection data have already proved important for systems biology, two great problems accompany these data, namely that coverage for *H. sapiens* is low, and that the experimental data are full of noise [276, 14]. Therefore we wish to construct methods to computationally validate experimental PPI data, as well as predict novel interactions. With this objective, we look at individual protein pairs, at their chemical properties, predicted functionality and domains. In chapter 4 we look at the network as a whole, and validate individual links based on the network neighborhood and topology.

In addition to the genetic/knowledge-based methods described above, there have been numerous efforts made to predict PPIs from primary protein sequence. This section is concerned with PPI prediction from sequence and sequence-derived features using machine learning and statistical methods.

Since its birth in 1993, the Center for Biological Sequence Analysis has had a strong profile in sequence annotation derived by machine learning and prediction. Prediction methods such as SignalP [24], and NetStart [209], to name two popular servers, can produce results on novel data in negligible time and therefore guide experimental work , thus saving time and labor. For the duration of my Ph.D. work at CBS I have striven to construct a prediction method that, for a given pair of protein sequences, can produce a qualified prediction whether they interact or not. This project turned out to be a much more difficult task than expected, as far as data preparation was concerned, mostly because the data are different than those generally used in sequence analysis in that a pair of sequences is under observation at a time. This fact renders many conventional approaches to sequence encoding and redundancy reduction useless.

In addition to describing our work, the following sections give an introduction to the machine learning algorithms employed in our work and highlight some of the difficulties involved in the prediction of PPIs.

3.2.1 Machine learning

Several types of machine learning methods have been employed to predict PPIs. Here I will briefly describe two machine learning techniques, commonly used for sequence-based purposes at CBS.

Artificial neural networks

Modeled after the nervous systems, artificial neural networks (ANNs) are complex structures of simple units called neurons, each having input and output signals. ANNs exist in a number of architectures and employ several different learning algorithms. They have been successful in many fields, including speech recognition, credit card fraud detection and model airplane auto-piloting. They are known for their general learning ability and robustness to noisy data.

Just like real neurons, the artificial neurons receive a number of input signals, each carrying an individual weight, a non-linear function, commonly a sigmoid, is applied to the weighted sum of inputs. If the result exceeds a predetermined threshold, an output signal is triggered; the neuron “fires”.

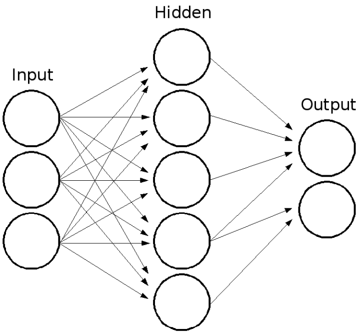


Figure 3.3: **A simple diagram of a 3-layered artificial neural network. The input layer scans and applies weights to the input vector. The hidden layer adds weights to the output from the first layer and finally, the output layer aggregates the signals from the previous layer and, depending on the aggregate and the firing threshold, either fires or not to make its binary classification.**

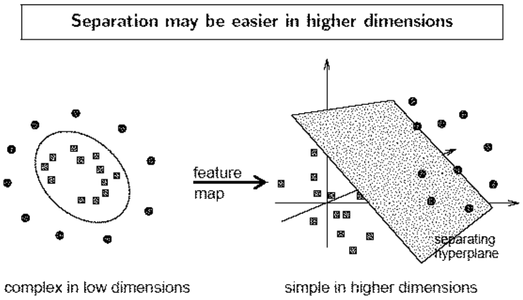

Separation may be easier in higher dimensions
feature
map
separating
hyperplane
complex in low dimensions
simple in higher dimensions

Figure 3.4: **The kernel trick. Support vector machines map data to a higher dimension, where they are better separable by a plane.**

These simple neurons are then arranged in complex, highly interconnected networks. The broad range of network architectures that has been described in the literature is beyond the scope of this thesis, readers are referred to [16] and references therein for a more general background of ANNs.

For the remainder of this short ANN introduction, we will focus on the network architecture employed in our work on PPI prediction.

We used a feed-forward type network, where the flow of information is unidirectional from the input layer, which in our case is where the sequence or sequence features are represented by vectors, to the output layer, where a classification is made by examining the values that the output classification neurons achieve. The PPI classification problem is binary: the pair of proteins under investigation either interacts or not, so there is only need for two output neurons, one for each case. Generally in an N-ary classifier, there is need for N output neurons.

The goal of the neural network is of course learning, which is obtained by training. This is in practice done by exposing the network to known classification examples, continuously updating the weights so that the output values match the known class. Our networks use the back-propagation [16] algorithm to update the weights.

Support vector machines

Support vector machines (SVMs) are a relative newcomer to the field of machine learning and classification. Although Vapnik described the basic idea of a hyperplane separation algorithm back in 1963, it was not until 1995 that such classifiers became practical and the name *support vector machines* was coined [272]. Since then, SVM applications have grown explosively in number and proved equal to or better than many other learning strategies in various classification tasks.

The basic idea behind SVMs is to map feature vectors from different classes to another "space" where vectors of different classes can be separated by a hyperplane. The mapping between spaces is performed by a kernel function, which can be one of several types. This so-called "kernel trick" allows a classifier to use linear separation in a higher dimension space to solve a problem, which maps back to a non-linear separation in the original problem space.

Of particular interest to our application of prediction PPIs are *linear, structural* SVMs [138], which can be trained in a much smaller time frame than most other classifiers and are therefore applicable in situations where the input is a vector of high dimensionality. Linear SVMs have proved successful in classification problems with sparse input vectors with several thousand dimensions [138].

Performance assessment

When performing classification and prediction tasks, the outcome is measured in true and false predictions. There are several composite metrics that make use of the true and false positives and negatives, the most common ones listed in table 3.1.

Specificity, or *precision*, is the ratio of true positive predictions to all positive predictions, while *sensitivity*, or *recall*, is the ratio of true positive predictions to all positive examples. There is usually a trade off between sensitivity and specificity for a given prediction method. A commonly used function to optimise the trade-off is the Matthew's correlation coefficient [185]. Another is the so-called receiver-operator characteristics (ROC) curve, which shows the specificity plotted against sensitivity. The area under the curve is a good measure of the performance of the classifier and thus a good function to optimise during training

[299]. For a more theoretic and detailed discussion on performance assessment in machine learning, readers are referred to [17].

Specificity	$\frac{tp}{tp+fp}$
Sensitivity	$\frac{tp}{tp+fn}$
Accuracy	$\frac{tp+tn}{tp+fp+tn+fn}$
Matthews correlation coefficient	$\frac{(tp\times tn)-(fp\times fn)}{\sqrt{(tp+fn)\times(tp+fp)\times(tn+fp)\times(tn+fn)}}$

Table 3.1: **Some common performance assessment metrics in machine learning.** tp, tn, fp, fn **stand for true positives, true negatives, false positives and false negatives, respectively.**

Many recently published methods on PPI prediction reach high precision and accuracy, but one must remember that interaction maps are extremely skewed data with the number of negatives (probably) outweighing the number of positives by a factor of hundred to thousand. In the case of the model organism budding yeast, *S. cerevisiae*, the latest estimate of the number of proteins is about 6000. Without considering constraining factors such as subcellular localisation, the possible number of pairwise interactions, including dimers/self-interactions, is $6000^2/2 = 18,000,000$. An estimate of the number of interactions in yeast is about 30,000 [104], so the ratio of interacting to non-interacting pairs is about 1/600 or 0,25%. Knowing this approximate ratio beforehand can lead to a simple prediction method: namely to predict one of every six hundred pairs to interact and the remaining 599 ones not to. Such a prediction method will produce 30,000 positive predictions and 17,970,000 negative ones when applied to the yeast proteome. By random hits this method will probably produce 50 true positives, 29,950 false positives, 29,950 false negatives and 17,940,050 true negatives. The number of “true” predictions is therefore 17,940,100, leading to accuracy of over 99%, which is of course highly misleading for such random guessing. This demonstrates the need to use unbiased performance metrics to validate a prediction method. See [17, 16] for more details.

Specificity	0.0017
Sensitivity	0.0017
Accuracy	0.9967
Matthews correlation coefficient	0

Table 3.2: **The performance metrics from table 3.1 when performing a random guess assignment of protein interactions in yeast. With a skewed data set such as the interactome, where the ratio of interacting to non-interacting pairs is** $1/600$**, one has to be careful in choosing meaningful performance metrics.**

3.3 Paper III

Manuscript in preparation

Perspectives on protein-protein interaction prediction

Páll Ísólfur Ólason[1]*, Thomas Sicheritz-Pontén[1]*, Lars Juhl Jensen[2] and Søren Brunak[1].

*** These authors contributed equally**

[1] Center for Biological Sequence Analysis, BioCentum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark.
[2] European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg Germany.

Abstract

Protein-protein interaction data describe connections between proteins in the cell, providing clues about their properties and functionality. Experimental interaction data are notoriously error prone and human data scarce. Therefore many labs have produced methods of predicting protein protein interactions, attempting to get a clearer picture of the interactome. Accurate protein-protein interaction predictions allow us to clean up these noisy data sets and provide interactome coverage for *H. sapiens.* There are, however, serious obstacles to successful interaction prediction and reasons to doubt that the task is as straightforward as recent literature indicates.

Introduction

Accurate prediction of protein-protein interactions (PPIs) has a potential to be a valuable tool for the future development of systems biology and for many other tasks within bioinformatics in general. Reliable predictions of interaction partners for any protein in a given pathway or complex can cut down time and cost in the laboratory and focus the experimental resources on targeted cellular studies of interest. Unfortunately, successful PPI prediction is not as straightforward as recent literature has indicated.

Comparing existing methods for PPI prediction is not trivial, as several types of data and methods have been utilised for this task. The earliest computational methods predicted the site of interaction in protein complexes (the "docking problem"), rather than whether two proteins interact or not. Methods for predicting the interaction sites in proteins are usually focused on the surface features of proteins of known structure [140], or on conserved sequence patterns in families of interacting proteins [45], or both [87]. Pazos and Valencia suggest that correlated mutations and the *in silico* two-hybrid method [206, 208] may be useful in the latter approach [267]. A review of docking prediction methods can be found in [241]. CAPRI (A Critical Assessment of PRedicted Interactions) [129], is a contest, that focuses on prediction of interaction sites of proteins to aid structural modeling.

With the emergence of high-throughput PPI detection methods [89, 119, 300], the interactome knowledge coverage has been vastly improved, even though the currently available data seem to suffer from low accuracy [276]. These large amounts of data allow for implementing machine learning methods in the field of predicting PPIs. Since the publication of high-throughput PPI data sets, several machine learning methods have been developed. Bock and Gough developed a data-driven prediction method, using calculated property vectors and support vector machines (SVMs) [32]. Martin et al. approached the problem in a similar way, but used a more sophisticated SVM kernel and a different type of input vectors, based purely on sequence [179].

The notion that protein domains or motifs, such as SH3 domains or coiled-coil structures, are the fundamental unit controlling PPIs has led to many interesting observations. Sprinzak and Margalit showed that there is a clear correlation between the domain composition of a protein and its interactions [247]. Deng et al. [67], Gomez et al. [103], Kim et al. [150] and Wojcik and Schäcter [286] have all constructed statistical methods, using domain information to predict PPIs at a better-than-random level. Domain information has also been used to train SVMs [71]. Betel et al. constructed networks of interacting domains and mapped those onto known biological complexes and pathways [28].

Recently, Jansen et al. integrated both high-throughput interaction data and indirect interaction evidence, such as co-expression, cellular function and essentiality, and created a Bayesian network for the prediction of interactions [132]. During this work a data set, named the yeast "gold-standard", was created, consisting of hand curated complexes from the Munich Information Center for Protein Sequences (MIPS) [187]. This data set has served as a benchmarking set in other work [179]. A similar approach, but aimed toward prediction of human PPIs, has also been published [220].

Data preparation in PPI prediction is a difficult task for two main reasons. First: instead of performing analysis of a set of single sequences, one must now focus on a pair of sequences at a time. This makes all data handling, such as database searches, sorting and filtering much more complex, as well as rendering standard methods of redundancy reduction and balancing useless. Second: there is no available set of unbiased, decidedly negative data; pairs of proteins that do not interact when coexisting in time and space. Such negative data is crucial for the training of most types of machine learning algorithms.

While the first-mentioned problem is quite an obstacle, it can be overcome with computing power and new algorithms. The second problem is much greater and a large part of our work described here deals with the definition and handling of negative interaction data and also how improper selection of negative training data may have affected previous attempts at PPI prediction.

Results

Pitfalls of previous attempts at negative data definition

As pointed out in the introduction, several groups have attempted to create methods to predict PPIs. The lack of negative training data and the high error rate of positive data complicates the task and there is reason to question many claims for high performance PPI prediction methods published as pointed out by Ben-Hur and Noble [22] and the following sections.

Combinatorial effect of large complexes

When interpreting complex pull-down data, using the "matrix" representation can result in a combinatorial explosion of pairwise interactions derived from large complexes.

Some of the previous attempts [132, 179, 175] have made use of the MIPS complex tables [187]. While these data are manually curated and of high quality, the combinatorial effect of the matrix representation of large complexes inherently biases the data. The single, largest complex, the ribosome, with almost 10,000 possible PPIs among its ~140 proteins, using the matrix representation, accounts for ~60% of the interactions in the "gold standard" set compiled by Jansen et al. Ribosomal proteins have specific features and properties and are conserved in sequence [96], so when using such data as input for a prediction method, along with functional annotations, sequence information and/or information of functional domains, there is a risk that the methods "learns" to recognise ribosomal pairs of proteins and in effect, becomes a predictor of ribosomal protein pairs, rather than a predictor of interacting pairs.

Correlation of sub-cellular localisation, domain composition and biological process

Several recent machine learning approaches to PPI prediction, mentioned above [132, 220, 171] pair together proteins from different sub-cellular compartments (e.g. a plasma membrane protein and a nuclear protein) as non-interacting pairs for training. While this assignment is presumably correct in most cases, it is inherently "dangerous" as input to a machine learning method because sub-cellular location is correlated to many kinds of biological information, such as amino acid composition [47, 49], protein domain composition [192, 41] and functional categories [50]. In the approaches mentioned above, information of sequence and/or functional categories was included as training data, and because of the correlation between sub-cellular location and sequence/function, it is quite possible that the method is biased towards sub-cellular location prediction rather than PPI prediction.

Using a similar approach as Rhodes et al. [220], we determined the likelihood ratio (LR) of a protein pair being in the positive set vs. negative set given the size of the so-called smallest shared biological process (SSBP), which is the entity farthest down the hierarchical Gene Ontology (GO) tree - in other words, the most specific term - which the two proteins have in common. As positive data we defined all human protein pairs from UniProt assigned by GO to either the plasma membrane or the nucleus. We then identified the SSBP of each pair and counted the frequencies of each SSBP size. The results are shown in table 3.3, and show a clear correlation of the SSBP size with the compartment assignment. This tells us that physically separated proteins tend to have a larger SSBP and therefore the negatively defined data Rhodes et al. use are already biased towards larger SSBPs, tainting this method as a neutral measure of a protein pairs likelihood to interact.

SSBP	GSP	GSN	Pr(S|GSP)	Pr(S|GSN)	LR
<10	2,498	76	0.00046	0.00149	0.31094
10-50	20,465	951	0.00379	0.01860	0.20358
50-100	39,984	1,874	0.00740	0.03664	0.20185
100-500	199,367	11,405	0.03688	0.22301	0.16537
500-1000	96,425	11,376	0.01784	0.22244	0.08019
possible	5,405,989	51,142			

Table 3.3: **Correlation of likelyhood of a protein pair being assigned to the same compartment and the size of the smallest shared biological process in Gene Ontology. As can be seen by the LR, it is becomes decreasingly likely for a protein pair to be assigned to the same compartment as the size of the SSBP becomes greater.**

We also calculated the so-called domain enrichment ratio for all the protein pairs in assigned to the same compartment (nucleus or plasma membrane) vs. all pairs where one protein was assigned to the plasma membrane and the other to the nucleus. We used 2/3 of the data to calculate the domain enrichment ratio, and the remaining 1/3 to test the distribution of enrichment ratios for protein pairs from the same and different compartments, respectively. While we did not see a clear correlation between the domain composition and the domain enrichment ratio on the whole scale of enrichment ratios obtained, only pairs assigned to the same compartment were seen having an enrichment ratio above 30, showing that protein domain pairs seen most commonly occur only in proteins assigned to the same compartment. All in all, this shows that there is a clear correlation between the input features and the negative examples in the approach taken by Rhodes et al. , Jansen et al. and others which make use of physically separated proteins as negatives for a feature driven classifier.

Random negatives

Yet another possible pitfall in the prediction of PPIs is the definition of negative data. Non-interacting pairs of proteins are hard to define and several approaches have been tried in this respect. The simplest approach is to simply define all pairs not known to interact as non-interacting [71]. This is a rather vague definition because of the non-completeness of the available PPI data. Bock and Gough used shuffled amino acid sequences as negative pairs [32] with good results. However, Lo et al. have assessed the effect of using such artificial protein sequences instead of real proteins in machine learning approaches [171], and found that using shuffled sequences gives a significant boost in perceived prediction performance, presumably because the algorithm learns to distinct between real proteins and artificial ones. It should therefore be avoided to use such shuffled sequences as input to a prediction algorithm.

Feature-based prediction

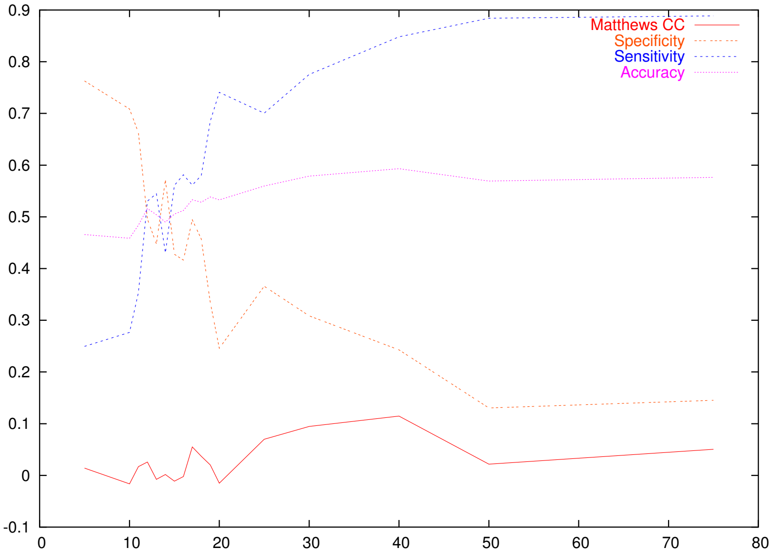


Figure 3.5: **Sensitivity, specificity, accuracy and Matthews correlation coefficient plotted against the number of times proteins were allowed to appear in the negative set of interacting proteins.**

Our initial approach using the ProtFun method [135], where calculated, derived and predicted protein sequence features are combined as input to a neural network simulator, showed great promise. We quickly achieved performance on par with previously reported figures using cross-validation, but when running the method on unseen data, the performance dropped to random, with almost all previously unseen protein pairs being predicted as interacting. A subsequent investigation revealed that the method had simply learned to recognise several proteins appearing multiple times in the negative set, while predicting all pairs of other proteins as positive. The graph-theoretical approach to defining non-interacting proteins (see section 3.3) had a side-effect of assigning isolated, peripheral proteins multiple times to the negative set.

As figure 3.5 shows, the specificity falls, and sensitivity grows as we allow the same protein to appear in more negative pairs. This indicates a rise in the false positive prediction rate and a drop in the false negative prediction rate caused by the ANNs simply "remembering" the most common proteins in the non-interacting set and predicting all subsequent pairs including those proteins as non-interacting and all other pairs as interacting.

We proceeded to construct an algorithm limiting the number of times a given protein was allowed to appear in the training sets. In order to loose as little data as possible, instead of setting a hard threshold, we created an algorithm to balance the ratio of occurrences of each protein in positive and negative data set. The algorithm uses the absolute value of the logarithm of the ratio of occurrences of each protein in the training sets as a target function to minimise. Each round in the algorithm prunes out a single interaction of the protein with the highest value of the target function and its interaction partner with the highest count of occurrences in the same set. A flowchart of this algorithm is seen in figure 3.6.

Using this balancing algorithm, we reconstructed our data sets, keeping the absolute log-ratio of occurrences of proteins within the positive and negative sets below 1.2. When we then proceeded to train the ANNs, using these redundancy-reduced data sets, the predictive performance fell dramatically, and we abandoned the feature-based approach as, apparently, the ProtFun type features do not capture the property differences of interacting vs. non-interacting protein pairs. Having said that, the most important features for PPI prediction, according to the feature selection algorithm, are sensible according to published material on protein-protein interactions. The are: *flexibility, disulfide bridge count, hydrophobicity, secondary structure* and *tyrosine kinase motifs.* These features were automatically selected in almost all runs of our training and therefore are likely to be important to the ability of two proteins to interact. Chain flexibility (and low count of cysteine bridges, which impose rigidity) was seen to promote interaction, which is not surprising. Schlessinger and Rost have shown that flexible regions, with no regular secondary structure are overrepresented in promiscuously interacting proteins [233]. Hydrophobicity is not surprising either, and the first analyses of PPIs focused on hydrophobic patch analysis [139].

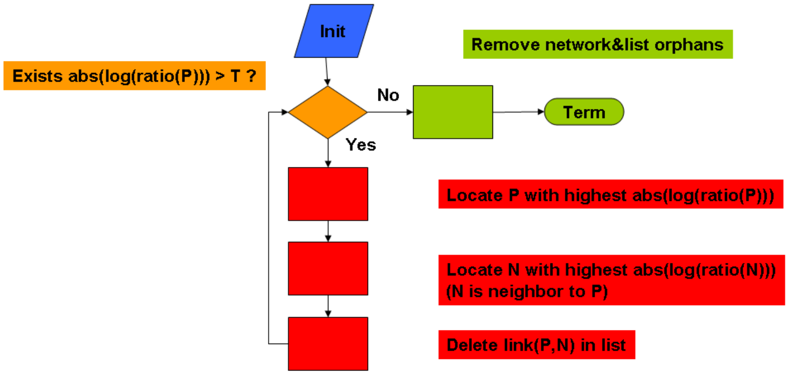

Init
Remove network&list orphans
Exists abs(log(ratio(P))) > T ?
No
Term
Yes
Locate P with highest abs(log(ratio(P)))
Locate N with highest abs(log(ratio(N)))
(N is neighbor to P)
Delete link(P,N) in list

Figure 3.6: **The balancing algorithm. A flowchart of the algorithm, used to balance the number of occurrences of each protein in the positive and negative training sets. The initial step is to count the frequency of each protein in the data sets. The ratio of occurrences in the two datasets is calculated and then the absolute value of the logarithm of this ratio is found as the target value. The protein with the greatest target value is identified and its interaction with the protein, having the second- highest target function value, is pruned from the database. This procedure is repeated until the ratio of all protein occurrences in the positive and negative training sets is within a given threshold.**

Domain profile approach

As discussed in the introduction, domain profiles have been used to infer interactions between protein pairs. The most common method is to analyse the frequency of domain occurrences in known interacting pairs and based on those frequencies, derive the odds of interaction between other pairs. While the procedure is simple and useful, and no non-interacting pairs are needed, it is limited in the way that large databases of PPIs are needed to derive accurate probability scores. When analysing interactions between multi-domain proteins it is also hard to state which domains are responsible for the interaction between the proteins. Therefore a "smarter" way of using domains as input for a PPI prediction method is preferable.

An obstacle is the size of domain vocabularies - the largest domain assignment databases, PFAM and Interpro [20, 193] having several thousand domain definitions. Such a vocabulary is prohibitively large for machine learning methods such as neural networks (ANNs) and support vector machines (SVMs) and results in long training times and complex network structure and requires large training data sets.

Recently, a novel training algorithm for SVMs was published [138], which allows training of a prediction method in time which grows linearly with the number of non-zero dimensions in the input vector. This is ideal for domain interaction vectors, which have a high dimensionality (the total count of domain definitions squared), yet they are very sparse (only a few of these vectors are non-zero - the product of the number of distinct domains in the two proteins).

We proceeded to create such domain interaction vectors, figure 3.7 shows the procedure of encoding the domain interaction matrix for a pair of interacting proteins. Using 10-fold cross validation, we reached a Matthews correlation coefficient of 0.5387, comparable to the training results of the feature-based approach discussed earlier. However, independent validation in the form of the yeast complex data from Gavin et al. [97] led to much higher scores than the feature-based ANNs. The results are seen in table 3.4.

At first inspection the method's performance on negative data seemed good, with 10 correct negative predictions for every false negative. The results on positive, however, data were not as good, as only 1 out of 8 true interactions is captured, reducing the value of this approach as a de novo predictor of PPIs. Furthermore a lack of domain definitions for about two thirds of the validation

Data	Specificity	Sensitivity	Accuracy	Matthews CC
Cross validation	0.8735	0.5040	0.7838	0.5387
Cellzome	0.2669	0.7699	0.7325	0.3367

Table 3.4: **Results from our domain-based SVM approach, both cross validation on our balanced training set and independent predictions on the data from Gavin et al. [97].**

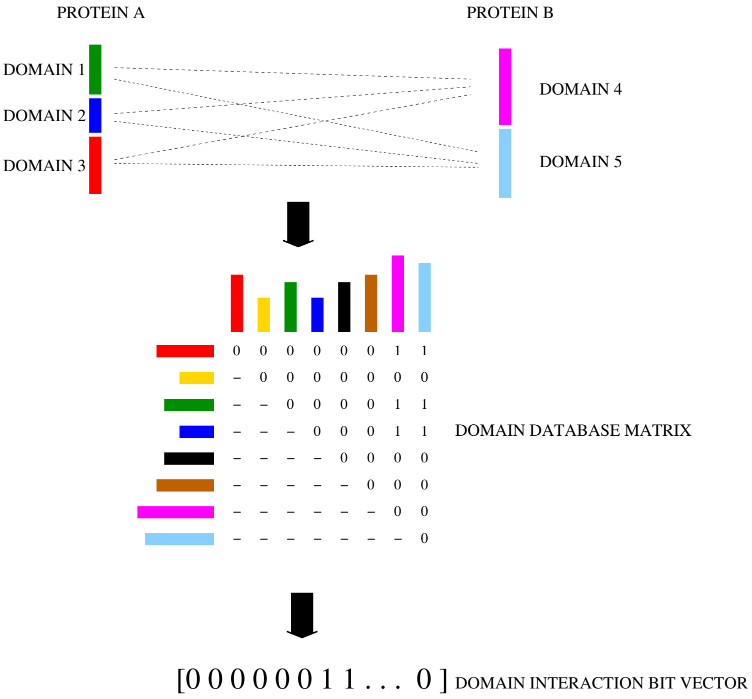
data set renders the majority of data useless in this case. Domain profiles for interacting pairs of proteins obviously hold predictive information about interactions, but result in low yield.

Discussion and future work

Data handling is tricky when working with pairs of sequences instead of the usual singles. Database queries, homology reduction and vector encoding for machine learning algorithm input are cumbersome and complex. We have identified a number of possible pitfalls in the handling of PPI data we believe have influenced previous results on PPI prediction. We produced a training set of unbiased data and defined non-interacting pairs of proteins with the help of graph theory. We constructed a novel balancing algorithm to homology reduce the positive and negative training sets maximising the amount of remaining pairs and we finally assembled prediction methods based on protein features using neural networks and domain composition using support vector machines. The feature-based PPI prediction approach yielded no useful results, while the domain-based approach seems more appropriate, but is hampered by low specificity on validation data, assigning most protein pairs as non-interacting. There are numerous data types which could be added to the PPI prediction model, such as sub-cellular location, functional category assignment, genetic evidence such as co-evolution and homology. Several previously published methods have made use of such data [132, 220]. In our search for a sequence or sequence-motif-based method to predict PPIs, we have refrained from using such data. In fact we have rigorously filtered our training data to exclude any sort of bias, due to sub-cellular location incompatibility, multiple occurrences of individual proteins/complexes etc. as we feel a sequence-based approach should. Our results indicate that protein features, which have proved successful in describing protein functionality in previous research [135] do not capture the essence of protein-protein interactions. Our domain-based effort were a bit more successful, yet not very useful as lack of domain definitions render this approach useful in only a third of the cases, and even then, the method is clearly biased to predict most pairs as non-interacting.

The sequence derived features used in the ProtFun approach do not seem to capture the nature of interacting proteins. This may not come as a surprise as protein binding is surface- and three-dimensional-structure-specific, and sequence-based methods have never been successfully applied to the problem of predicting protein structure. Protein domain definition usually carry structural information, and in that information the success of domain-based efforts may lie.

Many groups have attempted prediction of PPIs in the last few years, with varying results. It is worrying that the benchmarking data used in many of those attempts are flawed and as Ben-Hur and Noble conclude: "...prediction of protein-protein interactions from sequence is a difficult problem that can still be considered unsolved" [22]. A statement that we fully agree with as a result of our own experiences.



	🟥	🟨	🟩	🟦	⬛	🟫	🟪	🔹
🟥	0	0	0	0	0	0	1	1
🟨	–	0	0	0	0	0	0	0
🟩	–	–	0	0	0	0	1	1
🟦	–	–	–	0	0	0	1	1
⬛	–	–	–	–	0	0	0	0
🟫	–	–	–	–	–	0	0	0
🟪	–	–	–	–	–	–	0	0
🔹	–	–	–	–	–	–	–	0

[0 0 0 0 0 0 1 1 . . . 0] DOMAIN INTERACTION BIT VECTOR

Figure 3.7: **Encoding of domain interactions to a sparse vector for SVM training. For a given pair of interacting proteins (or non-interacting in the case of the negative data set), A and B, we scan the proteins for domains and create a matrix of all possible domain interactions (all domain database entries included). We then "flatten" that matrix out to a bit vector, which is the input for the SVM training.**

Materials and methods

Positive data

We used integrated data from the STRING database, taking advantage of the fact that STRING has implemented a quality score for PPIs and that STRING is an integrated resource, where there should not be a bias towards any one evidence type, organism or experiment. To set out with a high-quality data set, we chose to use only interactions with a STRING score above 0.8. We also chose to limit ourselves to budding yeast as a model, hoping to later expand our research to other, less studied organisms (PPI-wise) such as man. This initial data set contained 7324 interactions from complex pulldown experiments, using the spoke interpretation of experimental data.

Negative data

We tried to avoid pitfalls such as the ones described above in order to produce a negative data set of non-interacting proteins, not different from our positive data pairs in any way, such as complex-specificity, compartment bias or biological genuinity of the sequence. We turned to graph theory for definition of negative examples, assembling all the positive interactions we could get our hands on into a graph, and then selecting those protein pairs which were furthest apart in the network. To prevent compartment bias we filtered those pairs out of the set where the two proteins were not annotated as having been experimentally seen in the same compartment. The network distance was set to a minimum of 4 edges, which yielded 41,271 protein pairs to work with as non-interacting pairs. After running our balancing algorithm with a maximum absolute log frequency ratio of 1.2, 5281 protein pairs remained in our negative training set. Obviously, this method will only work on a reasonably complete interaction graph, as new links in the network may shorten the distance between any two proteins and disqualify the pair according to our definition of non-interacting pairs. While we believe that the method is valid in budding yeast, for which there have been published on average 10-20 PPIs pr. protein, it may not translate well to other, less studied organisms.

Feature encoding and selection

We applied the ProtFun method of feature-based functional prediction for proteins [135]. Feature vectors were calculated/predicted for each protein in the data sets and then concatenated for each pair appearing in either the positive or negative data sets. Using all features as input to a learning algorithm does not necessarily produce the best biological discriminator. As described in [135], training an ANN on smaller subsets of feature combinations will most of the time result in much better predicting performance than an ANN which has been trained on all features (the complete feature vector). The idea is to reduce the input space to combinations of features which in concert contribute high discriminatory value. We applied both a heuristic and a genetic algorithm for feature selection and trained ANNs using an optimal input of 7 features. Each pair of proteins was presented twice to the networks switching the order of proteins between the runs.

For the domain profile-based SVM approach, we created a contingency matrix for every possible domain-domain interaction, given a pair of interacting proteins. This bit matrix was then flattened into a single, sparse vector and dimensions that were never seen to be non-zero were pruned out for performance reasons.

ANNs

We used feed-forward type networks using backpropagation to update the weights. Training was performed with threefold cross validation.

SVMs

We used the SVM*light* package including the latest linear, structural SVM training algorithm [138]. Training was performed with tenfold cross validation.

3.4 Whole proteome interaction maps

As discussed in this chapter, the low reliability of PPI data, especially high-throughput data, shows the potential benefits for prediction methods for PPIs. Several approaches have been taken in this respect, but method comparison is difficult and results are easy to tweak with improper benchmarks and biased input data, as well as hard to validate.

As more and more PPI data are assembled, a network structure emerges and network analysis is a whole science in itself. The next chapter includes discussion about graph-theoretical methods to validate and score PPI networks on both local and global scale.

Chapter 4

Networks

Let us now take a look at the structure of sequence components which emerges when integrating various data types, superimposed on the frame of protein-protein interaction networks.

Several complex structures, such as hyperlinks on the Internet, social contact networks, disease propagation and electrical circuits have been modeled using networks [64, 279, 4, 18]. The huge networks that have emerged from protein interaction data have been the subject of much research in the last years, ranging from simple guilt-by-association transfer of annotations between physically interacting proteins to complex large scale graph-theoretical studies.

In the framework of the *disease gene finding* group at CBS, a warehouse of protein-protein interaction data has been constructed, integrating interaction data from most or all of the large interaction databases mentioned in chapter 3 as well as derived interactions such as pathway neighbors in KEGG [144] and Reactome [141] as well as orthologous interactions between different species. This huge resource, called *Inweb*, contains over 300,000 interactions in humans. Similar data sets have also been produced for model organisms such as budding yeast, fruitfly and nematode. These huge networks help us identify local functional modules and proposed interconnections when searching for disease-causing genes, as explained in section 4.2. Analysing the their global structure from a systems point of view is also interesting.

In contrast with the reductionism applied when performing analysis and annotations of the protein and gene components of cells, the network analysis of the interactome tends to take a birds-eye view of the data, integrating more and more information and will hopefully at the end produce a model that can explain systems of biological entities working in unison: a complex, a pathway, a regulatory mechanism, the cell.

This chapter contains some theoretical background on network theory as well as network analysis of the human interactome, focusing on specific chromosomes in hope to find clues to the mechanisms that govern chromosomal diseases such as trisomies. The chapter is concluded with a research paper on disease gene prioritisation. The paper revolves heavily around two networks, the human interactome and the phenome, a network of phenotype descriptions, where links are made by text mining disease descriptions and finding term overlap.

4.1 Graph theory for the biologist

In chapter 2 we saw all sorts of annotations being attached to the components of genes and proteins. Similarity in any of these annotation spaces, such as the localisation, co-expression or regulatory levels can be viewed as links between the components, as can more direct physical links as direct protein-protein interactions or metabolic reactions. An assembly of the components and links from any of these spaces can thus be pictured as a network showing the step-wise connections between all components. Such networks are not only useful to visualise a large number of components and their interconnections, there is an entire mathematical discipline devoted to the analysis of networks: *graph theory* [69]. This section should give readers a bit of background on subject.

4.1.1 Terminology

A *network*, or *graph*, is a collection of nodes and edges. For our purposes a *node* represents a protein or gene and an *edge* represents a link between nodes, implying a connection on some level, be it a direct physical interaction or a similarity in phenotypic description when the two genes or proteins are dysfunctional or something third and completely different.

In this section I will try to explain the general terms in graph theory, that have been related to biology, in an understandable language, leaving the math out as far as possible. If mathematical definitions and notations are of interest, readers are referred to [69].

Directed vs. undirected edges

One of a network's most important properties is whether or not it is *directed*, meaning whether or not an edge from node A to node B implies a reciprocal edge from node B to node A. Directedness results from, e.g., temporal order or some other irreversible action between the components, making the step from A to B possible, but the step back from B to A impossible. An example of networks often modeled as directed graphs are metabolic networks, where metabolites are chemically and (practically) irreversibly modified. Undirected networks model components and links between them, where the step between A and B is equal to the step between B and A or that the edge between A and B is unrelated to temporal terms. A graph modeling physical interactions in a protein complex is an example of a network that would generally be undirected, as there is little sense in applying direction to the structural interface between the two proteins.

Node degree

A node's *degree*, or *connectivity*, is the number of edges connecting that node to other nodes. For a directed graph, the degree can be split into out-degree and in-degree, meaning the number of edges starting at the node and the number of edges terminating at the node, respectively. Nodes with a high degree are often termed *hubs*. Although there is no formal definition of how many connections a node must have to qualify as such, Jeong et al. contrasted proteins with ≥ 15 connections with those having ≤ 5 connections as hubs and non-hubs, respectively [137]. Ekman et al. defined hubs as proteins with ≥ 8 interactions and non-hubs as those with less than 4 interactions [78].

Path lengths

The *path length* between two nodes in a graph is simply the number of edges separating the nodes. In a given graph, there may be several paths possible between two nodes. The shortest path, or *geodesic*, between two nodes is the the path between those nodes (there can be more than one), containing fewest links. The *diameter*[1] of a network is the average of all the pairwise shortest paths in the network [137]. A node's *eccentricity* is the length of the longest path from that node to any other in the network.

Clustering coefficient

The *clustering coefficient* of a node measures how many connections there are between the node's neighbors as a ratio of the possible maximum number of connections. This is given by $C_n = \frac{2e}{k(k-1)}$, where n is the node under inspection, e is the number of edges observed between any of n's neighbors and k is the number of neighbors n has.

[1] In classical graph theory, the diameter is defined as the maximum of the shortest paths in the network [213].

4.1.2 From topology to biology: global topology

The last section introduced some of the concepts that are used when dealing with graphs. Most of these relate to individual nodes and edges in the graph. When analysing graphs on the size order of the entire human proteome and interactions between the protein components, it is useful to have measures that describe the network and subnetworks as a whole and not just the individual nodes and their links. Analysing the network structures that emerge when integrating omics data is one of the biggest and most interesting tasks in systems biology.

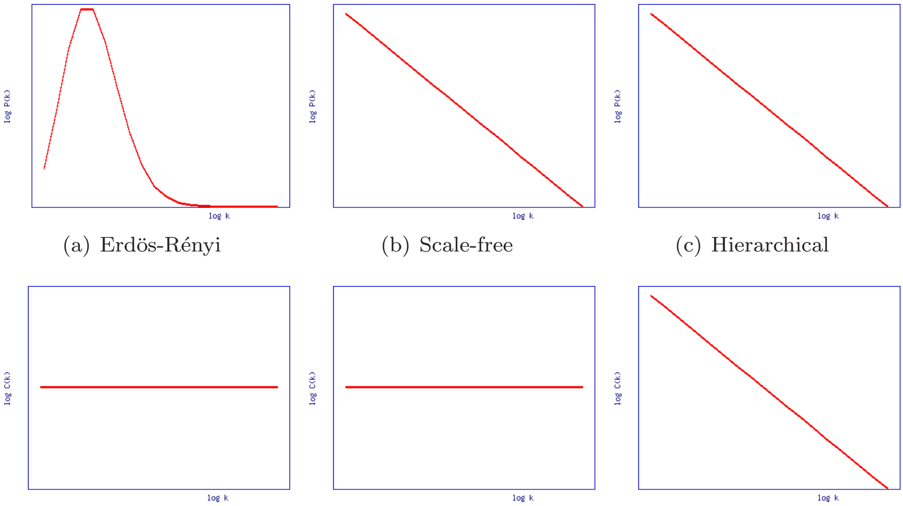
Network models and topological measures

Around 1960, Erdös and Rényi constructed a model for random networks where links are placed with equal probability between any two nodes [84]. As a result, the connectivity of the nodes in such networks follows a Poisson distribution (figure 4.1(a)). A property of these networks is that they are *small-world*, meaning that one can move from one node to any other by traversing only a minimal number of links. In mathematical terms the diameter of small-world networks is proportional to the logarithm of the network size.

Empirical observations of www hyperlinks, social contact networks, electrical circuits, and various biological networks, such as interaction networks, metabolic networks, and regulatory networks [18] have revealed that the global topology of these real-world networks differs from the Erdös-Rényi model, particularly, the connectivity distribution follows a power law: $P(k) \sim k^{-\gamma}, \gamma > 1$ (figure 4.1(b)). In order to quantify the statistical significance of the structure of many real world graphs, Barabási and Albert introduced a new graph model: the *scale-free* network [19], in which new edges preferentially attach to proteins with a high connectivity; hubs. It has been shown that many biological graphs, as well as the Internet hyperlinks and social networks better fit the scale-free model than the Erdös-Rényi model [137, 4, 18]. Scale-free networks are what is known as ultra-small-world, meaning that every two nodes are joined by a path even shorter than those in random networks; their diameter is proportional to $log(log(N))$, where N is the network node number[2] [52, 56].

[2] Actually, the $log(log(N))$ correlation is seen for power-law degree distributions where the exponent ranges from -2 to -3. Practically all scale-free networks inspected here qualify as such.

The new Barabási-Albert scale-free model has sparked much interest and a number of publications on networks. The concept of such heavy-tailed probability distributions is, however, by no means new [145]. Used to describe land ownership in Italy over 100 years ago by Pareto, such distributions were used as basis of Yule's mathematical models of "preferential attachment" in 1925 [294] and such "rich get richer" distributions were analytically described by Simon in 1955 [237] and later used to analyse scientist citation networks [64, 63].



(a) Erdös-Rényi (b) Scale-free (c) Hierarchical

(d) Erdös-Rényi (e) Scale-free (f) Hierarchical

Figure 4.1: **Upper row: degree distribution $P(k)$. For Erdös-Rényi networks, the degree probability follows a Poisson distribution, while for scale-free and hierarchical networks, the distribution follows a power law, indicated by a straight line with a slope of -1 on a log-log plot. Lower row: clustering coefficient $C(k)$ as a function of degree. The clustering coefficient is not correlated with node connectivity in Erdös-Rényi and scale-free networks. A slope of -1 on a log-log plot of $C(k)$ is an indication of hierarchical structure. Inspired by [18].**

Despite having a probability distribution of connectivity that favors hubs, a feature of PPI networks is that they contain a large number of clusters. The concept of heavily interconnected clusters may seem in direct opposition with the scale-free model, where few hubs direct the topology of the network. Yet, it has been shown that interaction maps, that display a degree distribution characteristic of scale-free networks, also contain functional modules [113]. Scale-free nature and modularity can be explained with *hierarchical modularity* [293, 217], where disjoint clusters form increasingly larger clusters. Hierarchical networks display scale-free properties, such as a power-law distribution of node connectivity, and at the same time the clustering coefficient in such networks shows a correlation with the node degree, which is not seen in regular Barabási-Albert scale-free networks (figure 4.1(f)) [293, 18].

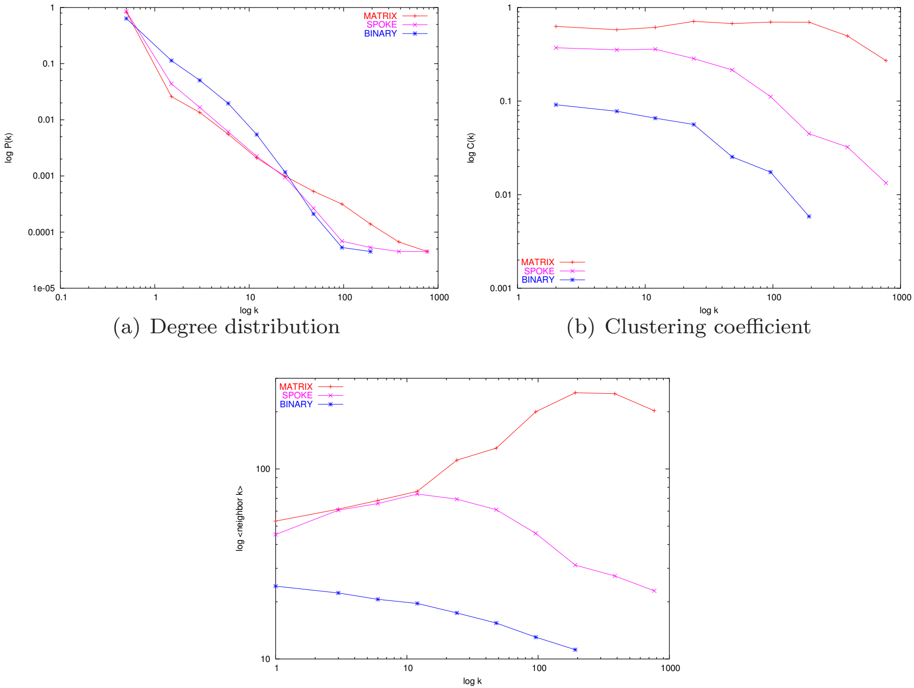
A recent publication argued that protein interaction networks are better modeled by geometric networks than the scale-free model [213]. Khanin and Wit statistically challenge the scale-free model as suitable for protein networks [148]. The geometric model implies that there are spatial constraints to the probability of two nodes being linked. While there undoubtedly are constraints to which proteins can interact, it is unclear how to best model these. The assumption that protein interaction networks are scale-free has also been challenged on the basis that the underlying data are incomplete and that Erdös-Rényi and other network models cannot yet be ruled out as fitting [109, 252]. Keller also criticises the global acceptance of scale-free distributions as some sort of global laws that governs most complex systems from our cells' proteins, to our sex lives to the Internet [145]. This is indeed an important point in today's network analysis of cellular data; the results so far are based on empirical observations and the data are known to be notoriously biased, incomplete and full of false positives [276, 134]. The dynamics, that underlie the properties displayed by these networks, will hopefully be better explained as they gain more experimental coverage.

Different experiments - different networks

The CBS data warehouse stores a large amount of protein-protein interaction data, both experimental and computationally derived. This resource, *Inweb*, is described later in section 4.2. The Inweb comprises 3 main data types: spoke, matrix and binary interactions which have been assigned to their categories by hand curation of experimental evidence. The spoke and matrix data are derived from complex pull-downs where pairwise interactions are derived from purified complexes. As described in chapter 3 there is a great difference in the resulting set of derived interactions; the spoke model may reduce the number of false positives, but may also miss a number of true interactions, whereas the matrix model includes all the possible interactions at the risk of producing many false positives. While the matrix model can be viewed as a continuation of the spoke model for hypothetical pull-downs of all the proteins in a given complex, a combinatorial problem arises for large complexes, as the number of interactions grows as the square of the number of proteins. E.g. a complex containing 100 proteins would give rise to 4950 matrix interactions, while the spoke model only yields 99. The so-called binary data are Y2H data and small-scale experiments where there is evidence of direct physical contact between the two proteins.

Figure 4.2 shows the results of our network analysis of the whole Inweb of human interactions. The degree distribution agrees with prior publications [101, 293] and displays scale-free properties as described in the previous section (a). The clustering coefficient shows correlation to node degree, which is indicative of a hierarchical structure in the network (b). This is most prominent in the binary interaction network. The third plot shows the average neighbor connectivity as a function of connectivity (c). Maslov and Sneppen have shown that yeast Y2H data has a declining slope on such a plot [180] which tells us that hubs tend to be disjoint and therefore the network is disassortative. The complex data tell a different story as both matrix and spoke interpretation of Inweb data show a rising tendency of hub interconnections until a connectivity of 10 and 100 is reached for spoke and matrix models, respectively, implying assortativity of hub proteins. In the case of matrix interpretation, this is understandable as large complexes will result in all the constituent proteins being interpreted as interconnected hubs, but seeing this effect in spoke data as well indicates that this is not just a combinatorial artifact, but a fundamental difference between networks constructed with CP data on one hand and Y2H data on the other.

This underlines the point that one must think carefully of what the network ,one uses to model complex systems, is meant to describe, and not to jump to conclusions from network analysis of incompatible data.



(a) Degree distribution

(b) Clustering coefficient

(c) Neighbor connectivity

Figure 4.2: **(a) Degree distribution in the human interactome network. All 3 data types: matrix, spoke and binary display degree distribution following a power law implying that they are scale-free. (b) Clustering coefficient ($C(k)$) as a function of connectivity. Correlation of $C(k)$ implies hierarchical structure of the network. All 3 networks show this tendency, but it is strongest in the binary network. (c) Average neighbor connectivity as a function of node connectivity for human protein interactions. Such data for Y2H experiments in budding yeast have previously been shown to follow a declining slope on a log-log plot [180]. This trend is also seen in our data for human binary interactions, which are mostly derived from Y2H experiments. This suggests that such networks are disassortative, meaning that highly connected proteins are less likely to be connected. For complex pull-downs the picture is quite different: both spoke and matrix interpretations yield a curve which has a maximum value, indicating that hubs have a tendency to connect to other hubs.**

4.1.3 Chromosomal interaction maps

This section describes work at CBS, where we have tried to use global topology measures, described in the last section, to quantify and analyse the interaction maps of individual chromosomes in the human genome. Ultimately our goal with this ongoing work is to identify discrepancies in the network structure of different chromosomes that may explain karyotype-phenotype relationship in chromosomal aberrations such as trisomies and some forms of cancer.

Chromosomes and phenotype

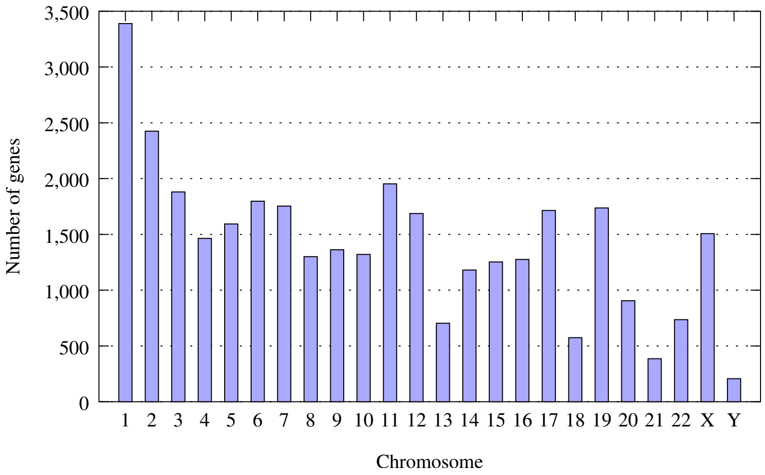


Figure 4.3: **Gene counts for all human chromosomes as annotated by the Ensembl genome browser (http://www.ensembl.org). The severity of developmental abnormalities of multiploidies 13, 18, 21 and of the Y chromosomes is correlated with the number of genes present on the chromosomes.**

Multiploidy, or polyploidy, is the genetically unnatural state of having more than the usual two copies (in the case of autosomes, one in the case of the sex chromosomes) of each chromosome in the cell´s nucleus, usually resulting from the failure of chromosomal segregation during cell division of gametes. By unknown mechanisms, this state is usually lethal, presumably because of an excessive copy number of the genes encoded on the chromosome present in multiple copies.

While most forms of polyploidy in humans are lethal, leading to spontaneous abortion, some are viable, namely the trisomies of chromosomes 13, 18, 21 and polyploidy of the sex chromosomes. Trisomy of chromosomes 13 and 18 is accompanied with severe defects and the affected individuals usually die shortly after birth but may reach their teens in some cases [21]. Trisomy 21, the most common cause of *Down syndrome*, has a complex phenotype of 71 described traits, including mental retardation, congenital heart defects, and a special formation of the face and hands [83]. Male X and Y diploidy and female triple X are known polyploidies, with mild phenotypes [223]. In the case of triple X, two Barr bodies[3] are formed instead of one. Males having the XXY constitution are usually tall, having androgynous features, and there are reports that this karyotype induces learning disability [223]. From 1965 to 1980, a number of studies were performed, linking XYY constitution with everything from tooth size to height to psychological problems and criminal behavior. Some evidence exists that the correlation with tooth size [7] and height [200, 151, 79] are real.

The most common common cause for aneuploidy is maternal meiotic non-disjunction [195]. The relative frequencies of non-disjunction are not random, but chromosome-specific with trisomy 16 as the most common form of aneuploidy. Maternal age is the only known risk factor in chromosomal non-disjunction, but the functional mechanism of this risk factor is unknown [195].

The viability and severity of the phenotype of the individuals suffering aneuploidy seems to be correlated with the number of genes on the chromosome with anomalous copy number, with a less severe condition being observed for chromosomes with a lower number of genes. Figure 4.3 shows the number of genes assigned to each human chromosome.

Aneuploidy and cancer have often been correlated and recently, a theory implicating cancer as a chromosomal disease, originally put forth over 100 years ago [110, 35], has been revived [75, 76]. If aneuploidy turns out to be a cause, rather than result, of some forms of cancer, a systems view of the chromosomes may aid cancer research and treatment in the future.

[3] In females, one of the X chromosomes is rendered inactive and forms a dense structure , known as a *Barr body*, peripherally in the nucleus.

Chromosome network topology

Using the Inweb network of experimental and derived protein interactions in humans, we extracted the subnetworks that correspond to single chromosomes. We then analysed the lists of chromosomal genes one at a time to see if any striking differences in network topology were observed.

Working from the hypothesis that genes are not essential by themselves, but rather take part in essential reactions and form vital complexes [115], we have looked at the interaction maps of individual chromosomes to see whether there is any evidence that the viability and severity of multiploidies could be related to topological effects in the interaction networks of these chromosomes.

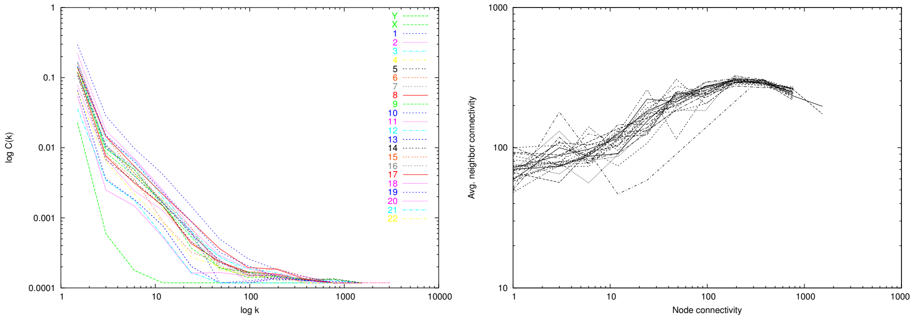


Figure 4.4: **Left: The degree distribution of the proteins on human chromosomes. While there are small differences, the slope on the curve is approximately the same, indicating that the degree distributions for the chromosomes follow power laws with similar exponents. Right: Average neighbor connectivity as a function of node connectivity. While previous reports indicated that neighbor connectivity showed a declining trend as the degree of the node increased [180], we find that the affinity of hubs to interact with hubs actually increases with hub size, up to degree ∼ 100. This trend is shown for the full interactome, as well as individual chromosomes and is reproduced with yeast and fruitfly data (not shown).**

Figure 4.4 shows the connectivity distribution of the human interactome, chromosome by chromosome. While the chromosomes with the lowest gene count also have the lowest connectivity, protein-wise, the slope of most of the lines is approximately the same, indicating that the chromosomal connectivity distribution follow the same power law.

Conservation of proteins

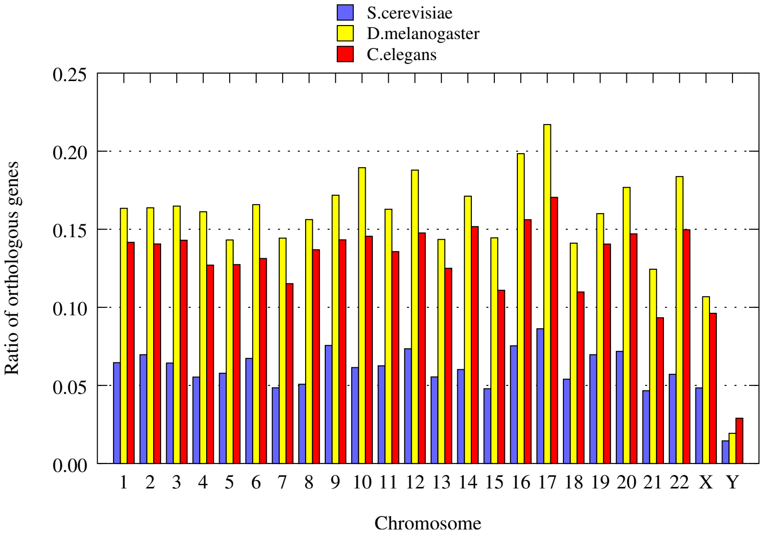


Figure 4.5: **Normalised number of orthologous genes grouped by the chromosome they map to in humans. The blue bars denote genes from budding yeast, the yellow ones fruitfly genes and the red ones are genes from the *C. elegans* nematode.**

Another hypothesis tested was that genes, conserved between species, are more likely to be essential than others and that the ratio of such conserved genes gives clues to the severity of chromosomal aberrations.

We assembled sets of orthologous proteins in budding yeast, fruitfly and nematode, and analysed which chromosome they mapped to in the human genome. Figure 4.5 shows the number of orthologous proteins normalised by the number of genes on the chromosome. The viable polyploidy chromosomes Y, 21, 18 and 13 do have a low ratio of orthologous proteins to these organisms, yet this is hardly conclusive evidence, as chromosomes 15 and 7 also show low ratios of orthologous genes to these species.

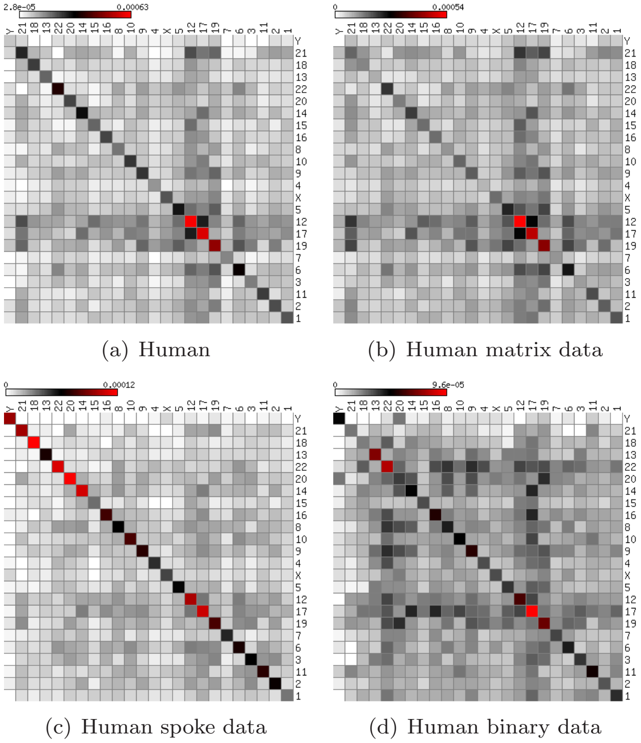
Physical proximity of genes encoding interacting proteins

In an attempt to see which chromosomes are most important for the human interactome, we mapped the chromosomal origin of all the PPIs in Inweb. Figure 4.6 shows the number of interactions between chromosomes, pairwise, normalised by the product of the two chromosomes' genes (except in the case of the same chromosome). The chromosomes are ordered by the number of genes residing on them to reveal, if any, size effects. Chromosomes 12, 17 and 19 have a higher than average count of interactions between them. Strikingly the fields along the diagonal are systematically darker than average, indicating that proteins are more likely to interact intrachromosomally, rather than interchromosomally. This effect is seen, using all PPI data combined, as well as by separating matrix, spoke and binary data types. We also created interaction networks for chimp, budding yeast, fruitfly and nematode as well to see if this effect could be reproduced in other organisms , and in most cases it is although the pattern is not so obvious in the case of the nematode as can be seen in figure 4.7.

The reason for this bias towards interchromosomal interactions is probably better co-regulation of interacting or functionally similar genes. Chromosomal colocation of functionally similar genes has been established before [55] and the same goes for genes that take part in the same pathway [167]. Co-expressed gene clusters have been shown to co-locate on chromosomes in mice [188] and there is a correlation between co-expression and interactions between proteins [99, 130, 132]. All these data and more were analysed by Teichmann and Veitia, who concluded that "genes encoding subunits of stable complexes are clustered on the yeast chromosomes" [258].

4.1.4 From topology to biology: local topology

After zooming out to view whole interactome maps, we now move from global topology to local topology. Most proteins work in concert with a number of other proteins to form a pathway or a complex [113, 164, 216, 121, 114, 236]. A great deal of research has focused on identifying such functional modules. Two main approaches have been followed, either focusing on biological descriptions of of proteins that are interconnected in interactomes or aiming at identifying statistically significant patterns in networks.



(a) Human

(b) Human matrix data

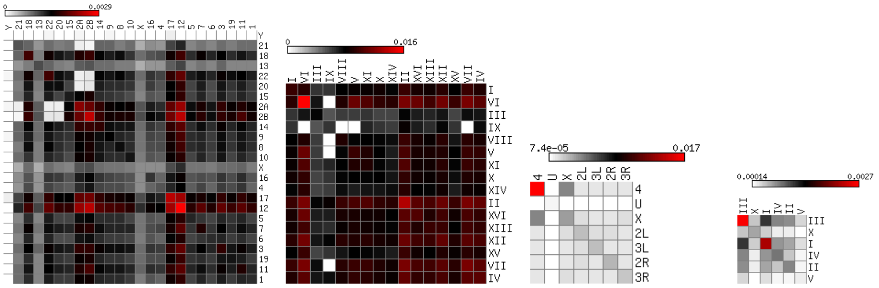
(c) Human spoke data

(d) Human binary data

Figure 4.6: **All-pairwise count of genes encoding interacting proteins, mapped to chromosomes and normalised by the total number of possible interactions. Note the dark fields along the diagonal, indicating that proteins are more likely to interact with other proteins from the same chromosome than they are to interact with proteins from other chromosomes. This striking effect is seen in all PPI data types.**

Centrality and essentiality

Centrality is a collection of terms in graph theory, relating to a node in the network. The centrality can be based on the degree of the node (degree centrality), the path length to other nodes (closeness centrality), its placement in the paths between other nodes (betweenness centrality), and other topological



(a) Chimp (b) Budding yeast (c) Fruitfly (d) Nematode

Figure 4.7: **All-pairwise count of genes encoding interacting proteins, mapped to chromosomes and normalised by the total number of possible interactions. From left: chimp, budding yeast, fruitfly and nematode worm. The diagonal effect, indicating that interactions are biased towards proteins encoded on the same chromosome is seen for all organisms.**

measures. The measures differ in directed and undirected networks, but the basic understanding of the terms is the same.

In general, the theory is that the greater a node's centrality, the more important it must be in the network. The scale-free nature of biological networks implies that there is an over-representation of hubs, nodes with a large number of interactions, which, by definition, have a high degree centrality. Such nodes have an important role in the graph structure, and hence also in the biological processes that underlie the links connecting the nodes. A hub in a metabolic network can be assumed to play multiple roles in metabolic pathways, examples of such hubs are multi-purpose kinases. Han et al. have identified two types of hubs in protein interaction networks: *date* and *party* hubs [108, 61]. Date hubs are named so because they interact with one partner at a time, temporally speaking, while party hubs are shown to interact simultaneously with many or all of their interaction partners.

Resulting from their topology, scale-free graphs have properties that are very robust to removal of random nodes, as these are most probably nodes with low connectivity. Targeted removal of the hubs, however, has devastating effects on the structure of the network [4], leading Jeong et al. and Han et al. to conclude that degree centrality and protein essentiality are correlated [137, 108]. He et al. have proposed that such central proteins should not be considered essential by themselves, but that they participate in more essential interactions than other proteins making them indispensable [115]. Sneppen and Maslov have recently found, that there is actually a negative correlation between essentiality and centrality in regulatory networks [181].

Topological scoring of protein-protein interactions

Realising that interacting proteins often cluster together, in the error-prone protein-protein interaction data, several labs have proposed topological scores to validate and clean these data [227, 228, 97]. We have used such methods to score our computationally derived network of protein-protein interactions 4.3.

Making use of the fact that interaction networks have a prevalence of hubs, Lappe and Holm have suggested to use a greedy algorithm to identify the nodes with a highest connectivity, not yet used as baits in interaction experiments [163]. They suggest that targeted experiments, using these highly connected nodes as baits, may provide 90% interactome coverage while only using 30% of the proteome as baits.

These algorithms rely upon knowledge of the network topology and obviously require as complete a network as possible for optimal function. As with many other properties of biological networks, they are based on empirical observations and benchmarks against known sets of interactions.

Motifs and statistical significance

As mentioned above, the models used to describe interactomes are based on observations and empirical data, and as such, constructed to understand the structure of real-world networks, rather than explaining them. Recently, these models have been used to assess the statistical significance of network structures. Such analysis is heavily dependent upon using a correct model and as discussed in the last section, a consensus has not been reached on the background distribution of interaction networks. Therefore, such statistical results must be viewed with skepticism.

Complete cellular networks are huge and in many cases we want to step down and identify smaller elements, or subnetworks[4] , commonly known as *motifs*. A classic motif example is a fully connected subnetwork - a *clique* - where all components are connected to each other. A protein complex with physical interactions between all subunits will be represented by such a structure in an interaction graph. Clustering algorithms aim at finding such completely or heavily connected subnetworks [82, 13].

An attempt to assess the statistical significance of a given motif or subgraph, a background model is needed. The scale-free model put forth by Barabási and Albert is usually considered the most proper model to date. A typical workflow to estimate the significance of a particular motif is to:

1. Select a motif from the network
2. Count the times the motif occurs in the network
3. Create a large number of random networks using the same topological model as the network being inspected (usually scale-free)
4. Count the times the motif occurs in the random networks
5. Assess whether the motif is significantly overrepresented in the original network compared to the random networks

Simple as it may seem, this is an enormous computational task. Selecting 3 nodes from a network of 1000 nodes yields well over 100 million possibilities, and going from those 1000 nodes to the whole human interactome with over 20,000 proteins, the possibilities grow enormously. Finding identical motifs in graphs and subgraphs is also a very complex problem. Several groups have tried to solve the problem using subset sampling and heuristics to minimise the computational time involved [189, 281].

Statistical results are only valid if an appropriate background model is chosen as a comparison. Usually the scale-free model is chosen, where the only constraint on new edges is that they are more probable to attach to hubs that already have a high number of edges. Real-world networks may be subject to more restriction that so. Scale-free networks have been used to model the neural connections in *C. elegans* and food webs and as a result several motifs were reported to

[4] A subnetwork is a substructure of a network, containing only nodes and links from the original network. Motifs are subnetworks that are significantly more common in a network than expected by random.

occur significantly more often in these networks than in corresponding random scale-free graphs [189]. Those publications do not take into account the spatial position of neurons, in the nematode nervous system or the placement of animals in food webs, as pointed out by Artzy-Randrup et al. in [9]. Yet it is intuitive that the farther apart in the system two neurons are, the less likely they are to be connected. The same goes for animals in ecological food webs.

The hierarchical clustering of PPI networks has been demonstrated in [217] and [101] and in the previous sections (4.1.2). Giot et al. proposed that this hierarchical nature was two-fold: the interconnections within clusters and then the network of interconnections between clusters. A similar viewpoint can be reached for the dynamic assembly of complexes. de Lichtenberg et al. [61] showed that complex formation is constrained in both space and time. Such constrains undoubtedly also apply to the higher level of cluster interconnections in the interactome, where physical barriers, such as cellular membranes, as well as co-expression compatibility must be acknowledged when performing large scale analysis.

Eukaryotic cells are physically divided into compartments, such as the nucleus, endoplasmic reticulum and mitochondria, by membranes. Such membranes compose a natural barrier to interactions between proteins, that reside in different compartments. As proteins are targeted to specific organelles, this compartmentalisation of proteins leads to a spatial bias in the interactome, which, as of yet, has not been accounted for in the background models for calculations of statistical significance of motifs. Figure 4.8 shows the inherent danger of modeling the interactome as an unconstrained network. A randomised version of a constrained network looks a lot less clustered and, when used as a null model, may artificially boost the significance of the motifs found in the original network. There are both spatial and temporal constraints on the possibilities of two proteins interacting and one must account for these to create more realistic null models when assessing subnetwork significance.

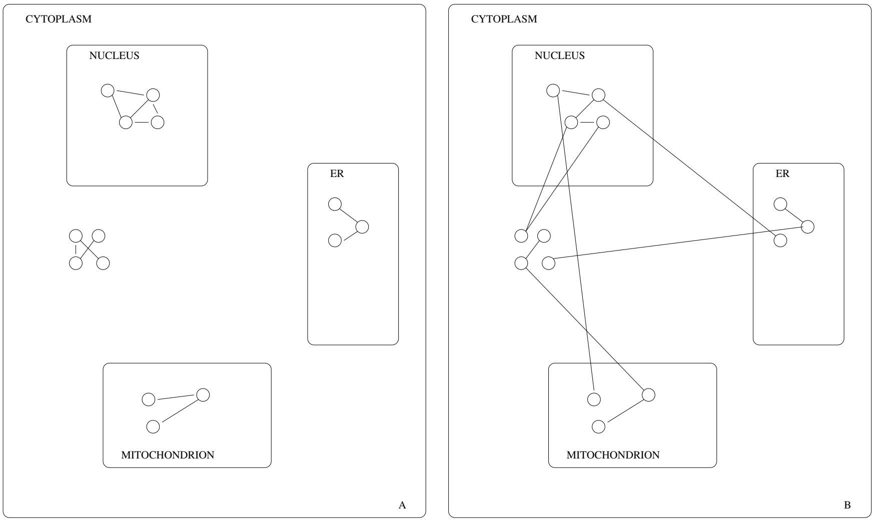


Figure 4.8: **These diagrams show a very simplified overview of a eukaryotic cell with 3 organelles: the nucleus, endoplasmic reticulum (ER), and mitochondrion. A membrane physically separates the organs from the cytoplasm. The small circles represent proteins and the lines interactions between those. Many proteins are targeted to a specific cellular compartment and thus, no interactions will occur between proteins in separate organs as seen in diagram A, on the left. On the right, the interactions have been randomised, to a point, while maintaining the connectivity of individual nodes. This is common procedure to make null models in scale-free networks [189, 281]. Comparing the diagrams shows the problems of such null models where the constraints of interactions, whether spatial or temporal, are not taken into consideration when randomising the networks. Compared to the incomplete null model on the right, the experimental network on the left displays a high level of clustering and therefore motif significance may be boosted.**

4.2 Disease gene finding: a case study in data integration

Performing network analysis on protein-protein interaction data may reveal subgraphs, or motifs, where the constituent proteins work in synergy in a functional module [289, 275, 113, 246]. Thus, functional information may be derived from protein-protein interaction maps via topological analysis. CBS' disease gene finding group has worked on the problem of assigning priority rank to candidate genes in hereditary human diseases for the last few years.

Working from the hypothesis that proteins function in a modular fashion and that a loss of functionality - stemming from mutation, incorrect folding or other problems - in any of a functional module's constituent proteins renders the whole module dysfunctional, resulting in similar disease phenotypes.

This following paper is the result of the work of the disease gene finding group at CBS. The paper naturally belongs at the end of this text, not only this chapter, but the whole thesis, because it is a showcase for almost all types of data integration and analysis discussed in this thesis. It represents a case study in biological data integration, data warehousing as well as federation of several types of annotations. Data formats as different as raw text and structured PSI-MI XML data are combined using machine learning and statistical approaches. Phenotype annotations are mapped to genes and proteins through text mining and a huge dictionary of gene and protein names and synonyms employed to expand our coverage of the human interactome as much as possible. The collection of annotations is glued on a scaffold of a protein-protein interaction network, composed of data from various sources, ranging from low-throughput, high-confidence experiments, to computationally inferred interactions from different species. The resulting phenome-interactome is inspected and validated using network analysis and a list of suspected disease genes is made available to the community.

At the end of all this data integration a framework emerges, which significantly improves the success rate of disease gene predicting compared to existing methods when using known disease genes for benchmarking.

4.3 Paper IV

Nature Biotechnology; 25, 309-316 (2007); doi:10.1038/nbt1295

A human phenome-interactome network of protein complexes in genetic disorders

Kasper Lage[1]*, E. Olof Karlberg[1]*, Zenia M. Størling[1], Páll Í. Ólason[1], Anders G. Pedersen[1], Olga Rigina[1], Anders M. Hinsby[1], Zeynep Tümer[2], Flemming Pociot[4,5], Niels Tommerup[2], Yves Moreau[3] and Søren Brunak[1].

*** These authors contributed equally**

[1] Center for Biological Sequence Analysis, BioCentum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark,
[2] Wilhelm Johannsen Centre for Functional Genome Research, Building 24.4, Blegdamsvej 3, DK-2200, Copenhagen N, Denmark,
[3] Department of Electrical Engineering (ESAT), Faculty of Engineering, Katholieke Universiteit Leuven, B-3001 Heverlee, Belgium
[4] Steno Diabetes Center, Niels Steensesvej 2, DK-2820 Gentofte, Denmark,
[5] Institute for Clinical Science, University of Lund, SE-22100 Lund, Sweden.

Abstract

We performed a systematic, large-scale analysis of human protein complexes comprising gene products implicated in many different categories of human disease to create a phenome-interactome network. This was done by integrating quality-controlled interactions of human proteins with a validated, computationally derived phenotype similarity score, permitting identification of previously unknown complexes likely to be associated with disease. Using a phenomic ranking of protein complexes linked to human disease, we developed a Bayesian predictor that in 298 of 669 linkage intervals correctly ranks the known disease-causing protein as the top candidate, and in 870 intervals with no identified disease-causing gene, provides novel candidates implicated in disorders such as retinitis pigmentosa, epithelial ovarian cancer, inflammatory bowel disease, amyotrophic lateral sclerosis, Alzheimer disease, type 2 diabetes and coronary heart disease. Our publicly available draft of protein complexes associated with pathology comprises 506 complexes, which reveal functional relationships between disease-promoting genes that will inform future experimentation.

Introduction

Several diseases with overlapping clinical manifestations are caused by mutations in different genes that are part of the same functional module. In such instances, the clinical overlap can be attributed to mutations in single genes rendering the complete module dysfunctional [39]. This concept has been applied to searches for disease genes by several computational methods, including, for example, schemes based on Gene Ontology annotations and gene expression data [2, 91, 92, 262, 210, 211, 182, 269, 270, 123, 93]. The advent of proteome-wide interaction screens in model organisms has revealed the modularity of the cellular interactome and that many genes exert their functions as components of protein complexes such as cellular machines, rigid structures, dynamic signaling or metabolic networks and post-translational modification systems [18].

Analyses involving model organisms, and more recently humans, show that direct and indirect interactions often occur between protein pairs responsible for similar phenotypes [2, 91, 92, 262, 210, 211, 182, 269, 270, 123, 93]. In humans this relationship can, for example, be observed in various inherited ataxias [170]. These findings hint at the widespread association of protein complexes with human disease and the likelihood that defects in several proteins, alone or in combination, can cause overlapping clinical manifestations. Systematic investigation of these complexes would help to elucidate cellular mechanisms underlying various disorders and prioritize positional candidates identified, for example, by linkage analysis or association studies.

Our strategy is predicated on the simple assumption that mutations in different members of a protein complex (predicted from protein-protein interaction data) lead to comparable phenotypes, the similarities of which can be automatically recognized by text mining. Computational integration of phenotypic data with a high-confidence interaction network of human proteins is required to perform such an analysis for many human diseases simultaneously. This creates a phenome-interactome network. However, there is no single standard vocabulary for phenotypic annotation in humans. Furthermore, protein interaction data are noisy, are scattered among different databases and contain many false positive interactions [276]. Additionally, only a few large-scale protein interaction studies have been finalized for the human proteome [225, 249] rendering the coverage of human protein interaction data too low for a systematic study of protein complexes associated with human disease. Thus, extensive data integration, including conservative incorporation of protein interaction data from model organisms, streamlining of human phenotype data and thorough testing of the resulting method, is required for the systematic investigation of protein complexes associated with human disease.

Results

Construction of a quality-controlled interaction network of human proteins and implementation of a thoroughly benchmarked computational phenotype similarity score allowed us to analyze a human phenome-interactome network. The results show that the 506 disease-associated protein complexes span a wide range of inherited disease categories. We furthermore trained a Bayesian predictor to prioritize candidates in 870 linkage intervals by assigning candidates to protein complexes and ranking these complexes based on the phenotypes associated with its members by text mining. The key steps in our approach are illustrated in Figure 1. Four disease-specific case studies are presented to illustrate how the complexes can be exploited to generate novel hypotheses, which directly suggest specific validation experiments involving particular patient-derived materials.

Measuring phenotype similarity scores

Text mining techniques are well suited for investigating phenotype-genotype relationships [182, 123, 93, 268, 155, 232, 40]. Inspired by such techniques, we created a scoring scheme that quantitatively measures the phenotypic overlap of Online Mendelian Inheritance in Man (OMIM) [106] records (Supplementary Fig. 1 online). For every record we created a phenotype vector consisting of weighted medical terms present in the record, which represent the phenotype described in that particular record. The parsing of the OMIM records was done using MetaMap Transfer (MMTx) [8], a program that maps text to the Unified Medical Language System (UMLS) [33] metathesaurus (MTH) concepts. The pairwise phenotypic overlap between records was quantified by calculating the cosine of the angle between normalized vector pairs [229], which is a standard measure in such analyses. Essentially, the method amounts to detecting words (from the UMLS vocabulary) that are (i) common to the description of the two phenotypes and (ii) do not occur too frequently among all phenotype descriptions and thus are informative about the phenotype under consideration.

Even though our approach is comparable to successful methods reported in other contexts [40], there are a number of problems surrounding the use of MMTx and UMLS [70], and it is not obvious that the cosine distance between phenotype vectors can accurately capture and quantify the phenotypic overlap between record pairs. To evaluate the reliability of our method, we extracted a large set of ˜7,000 OMIM record pairs, which had a high degree of phenotypic overlap. This assertion of phenotypic overlap was based on a combination of the opinion of expert OMIM curators and experts familiar with the diseases under consideration (Supplementary Methods online). To evaluate the phenotypic overlap of record pairs in this set, we manually curated 100 random record pairs. This evaluation showed that over 90% of the pairs consist of records with a high degree of phenotypic overlap (Supplementary Table 1 online).

The reliability of the phenotype similarity score was then tested by fitting a calibration curve of the score against the overlap with the OMIM record pairs (that is, the percentage of the pairs with a given score found among the record pairs). This demonstrates their direct correlation (Supplementary Fig. 2 online). The higher the phenotype similarity score between records measured by our text-mining scheme, the higher the probability that the records had been independently evaluated to have a phenotypic overlap by the OMIM curators, so that indeed the constructed phenotype vectors and scoring scheme produce a reliable measure of phenotypic overlap between OMIM records.

Constructing a scored network of human protein interactions

We created a human protein interaction network by pooling human interaction data from several of the largest databases and increased the coverage by transferring data from model organisms. We then devised and tested a network-wide confidence score for all interactions. This score relies on network topology and furthermore considers (i) that interactions from large-scale experiments generally contain more false positives than interactions from small-scale experiments [276], and (ii) that interactions are more reliable if they have been reproduced in more than one independent interaction experiment [276]. The reliability of this score as a measure of interaction confidence was confirmed by fitting a calibration curve of the score against overlap with a high-confidence set of about 35,000 human interactions (Supplementary Fig. 3 online). The resulting network contains ˜343,000 unique interactions between ˜8,500 human proteins. Of these, ˜62,000 are high-confidence interactions.

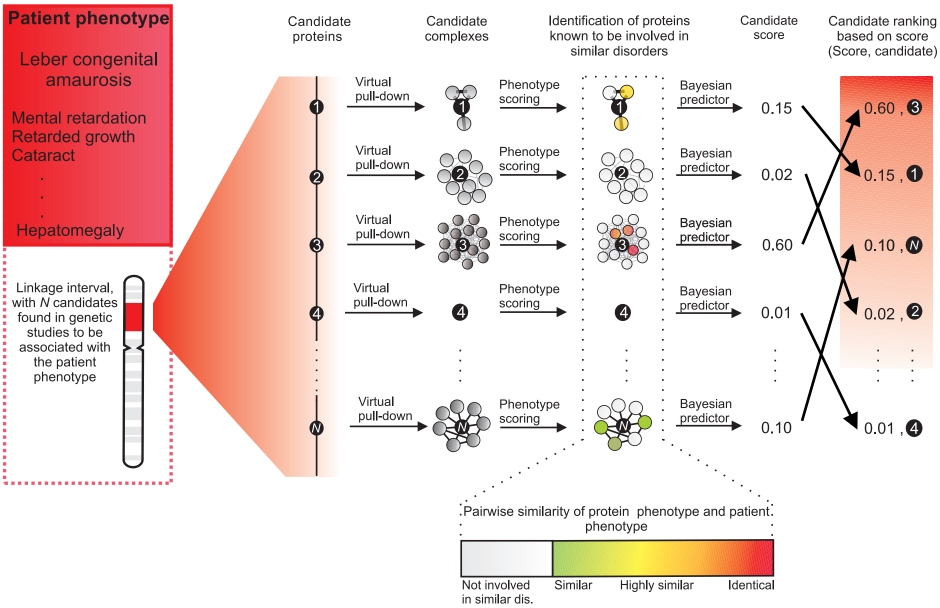
Testing the predictor on 1,404 linkage intervals

We trained a Bayesian predictor to rank known disease-causing proteins in linkage intervals, by assigning candidates to protein complexes and ranking these complexes based on the phenotypes assigned to their members by text mining. The predictor was validated by fivefold cross-validation on a total of 1,404 linkage intervals containing an average of 109 candidates and including one candidate known to be involved in the particular disease. For ranking candidates, the Bayesian predictor takes as input the patient phenotype (e.g., Leber congenital amaurosis) and a linkage interval, and the candidates are ranked by the following three steps (Fig. 1). First, a given positional candidate is queried for high-scoring interaction partners (termed a virtual pull-down of the protein). These interaction partners compose the candidate complex. Second, proteins known to be involved in disease are identified in the candidate complex, and pairwise scores of the phenotypic overlap between diseases of these proteins and the candidate phenotype are assigned. Third, based on the phenotypes represented in the candidate complex, the Bayesian predictor awards a posterior probability score to the candidate in the complex. All candidates in the linkage interval are ranked on the basis of this score. The biological interpretation of a high-scoring candidate is that this protein is likely to be involved in the molecular pathology of the disorder of interest, because it is part of a high-confidence candidate complex in which some proteins are known to be involved in highly similar (or identical) disorders.

Performance of the Bayesian model relying on phenomic scoring of protein complexes associated to disease

The results of prioritizing candidates in the 1,404 test linkage intervals show that the predictor has both good precision and recall (Fig. 2a). For each disease, we consider the known disease gene as the relevant gene. Our method makes a prediction for a disease if the top scoring gene for this disease has a

Lage et al. 2006
Figure 1


Patient phenotype
Leber congenital amaurosis
Mental retardation
Retarded growth
Cataract
Hepatomegaly
Linkage interval, with N candidates found in genetic studies to be associated with the patient phenotype
Candidate proteins
Candidate complexes
Identification of proteins known to be involved in similar disorders
Candidate score
Candidate ranking based on score (Score, candidate)
Virtual pull-down
Phenotype scoring
Bayesian predictor
1
2
3
4
N
0.15
0.02
0.60
0.01
0.10
0.60 , 3
0.15 , 1
0.10 , N
0.02 , 2
0.01 , 4
Pairwise similarity of protein phenotype and patient phenotype
Not involved in similar dis.
Similar
Highly similar
Identical

Steps in the scoring of each candidate in a linkage interval. 1) A virtual pull-down of each candidate is made identifying putative protein complexes including the candidate. This complex is named the candidate complex. 2) Proteins involved in disease are identified in the candidate complex, using the computational phenotype similarity score. In the figure proteins that are involved in different disorders comparable to Leber congenital amaurosis are colored according to the clinical overlap with this phenotype. 3) The last step is scoring and ranking of the candidates based on the complexes by the Bayesian predictor.

score above the threshold of 0.1. This threshold is chosen because predictions scoring below 0.1 approximate the random chance of picking the correct gene. The retrieved gene is then this top scoring gene. Precision (at a given threshold) is the proportion of relevant genes among all retrieved genes (# relevant genes retrieved / # genes retrieved). Recall is the fraction of the relevant genes that have been retrieved at the same threshold (# relevant genes retrieved / # relevant genes). For the 1,404 linkage intervals, there are 669 different predictions with a score above 0.1. Among these there were 298 correctly identified disease genes, so that the precision at this threshold is 45% (meaning that 45% of the candidates that ranked number one with a score above 0.1 are the correct disease gene) (Fig. 2a). This precision is much better than random. At this threshold the recall is 21%. A plot of precision versus prediction score cutoff shows proportionality between the score and the chance that the candidate is correct. Candidates scoring above 0.9 are correct in more than 65% of the cases (Fig. 2a). Thus, high scoring candidates are very likely to be correct, and the score awarded to a candidate is a direct indication of the chance that the gene is involved in the disease in question.

Performance of the Bayesian model relying on phenomic scoring of protein complexes associated with disease The results of prioritizing candidates in the 1,404 test linkage intervals show that the predictor has both good precision and recall (Fig. 2a). For each disease, we consider the known disease gene as the relevant gene. Our method makes a prediction for a disease if the top-scoring gene for this disease has a score above the threshold of 0.1. This threshold is chosen because predictions scoring below 0.1 approximate the chance of picking the correct gene randomly. The retrieved gene is then this top-scoring gene. Precision (at a given threshold) is the proportion of relevant genes among all retrieved genes (no. of relevant genes retrieved/no. of genes retrieved). Recall is the fraction of the relevant genes that have been retrieved at the same threshold (no. of relevant genes retrieved/no. of relevant genes). For the 1,404 linkage intervals, there are 669 different predictions with a score above 0.1. Among these, there were 298 correctly identified disease genes, so that the precision at this threshold is 45% (that is, 45% of the candidates that ranked number one with a score above 0.1 are correctly identified as genes causing disease) (Fig. 2a) - a level of precision far superior to random prediction. At this threshold, the recall is 21%. A plot of precision versus prediction score cutoff shows proportionality between the score and the chance that the candidate is correct. Candidates scoring above 0.9 are correct in more than 65% of the cases (Fig. 2a). Thus, high-scoring candidates are very likely to be correct, and the score awarded to a candidate is a direct indication of the chance that the gene contributes to the disease in question.

There were two main types of failures to identify the relevant genes. Either the proteins coded by the relevant genes do not have an interaction partner that is involved in a relevant phenotype (which applies to 59% of all intervals), or there is a gene in the region considered a better candidate by the predictor (which applies to 26% of all intervals). These 26% could in theory be correct predictions, as suggested by manual inspection of false predictions with high posterior probabilities. By far the most common failure is the lack of interaction partners involved in similar diseases. In 75% of such cases there were no candidates that scored above the threshold of 0.1. These failures could either be due to a lack of data or because some disease proteins do not interact with proteins involved in similar diseases. It seems most likely that the failures are due to a combination of both.

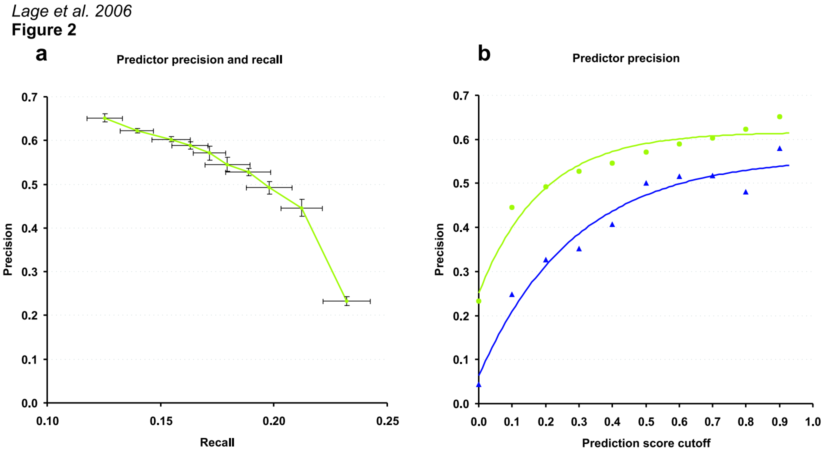
We also tested a predictor trained on large-scale protein interaction data from which bias related to human diseases was eliminated (Supplementary Methods online). Here we observed a comparable precision to the predictor trained on the full protein interaction data set (Fig. 2b). Using these data, the precision above 0.1 is 25%, and above 0.9, it is 58%. Therefore, although the performance is slightly lower, it is still very high. These results illustrate the value of large-scale protein interaction data from model organisms, if subjected to stringent quality control. The much lower recall (2.3%) is to be expected with less data. This shows that it is possible to accurately identify disease genes using data from model organisms that were not produced specifically to investigate disease relationships.

Because mutational analysis of candidates in linkage intervals is extremely demanding in terms of resources, our method should be valuable for identifying highly likely candidates and thereby facilitating the discovery of novel genes involved in human disease.

Predicting novel disease gene candidates

OMIM contains 870 intervals linked to diseases for which there are no confirmed disease-causing genes. We ranked the genes in these intervals by the method depicted in Figure 1. The full set of predictions above the threshold of 0.1 can be seen in the Supplementary Data. We present the best-scoring candidates made by our predictor in Supplementary Table 2 online. In each of the 91 represented intervals at least one candidate scores above 0.2. In some intervals there are also candidates scoring in the range 0.1-0.2, these are included for completeness, so the table contains a total of 113 candidates in 91 intervals.

All predictions in Supplementary Table 2 were followed up by independent literature studies, where we investigated the distance of the predicted gene to the closest published high-resolution marker. Seven genes were located >20 Mb from such markers (labeled * in Supplementary Table 2 online). We also investigated



Benchmarking results. Prediction scores are plotted against fraction of true positives (a+b). For predictors trained on both sets of protein interaction data, there is a comparable linear correlation between prediction score and fraction true positive predictions, showing that the score is a reliable measure of prediction accuracy. At scores > 0.9 the prediction accuracy is 65%. We also trained a Bayesian predictor on unbiased large scale data alone (b), with a resulting specificity comparable to the one using all protein interaction data. Thus, bias in the protein interaction data is not influencing specificity of the predictions.

whether the candidates had previously been associated with the respective disorders, and whether there were inconsistencies between candidates we proposed and those proposed by other groups for the same diseases and intervals.

Twenty-four of the predictions point to genes that are most likely true positives, but where the causative mutation has not yet been identified (annotated with ”2” or ”2#” in Supplementary Table 2 online). In these cases, our predictions should be seen as further evidence that the genes are involved in the respective diseases. Seven predictions point to genes where a causative mutation has been identified (annotated with ”3” in Supplementary Table 2 online). Together, these constitute 31 predictions most likely to be true. Of these, 25 are the best scoring in the interval, and 6 are scored second or lower. Sixteen predictions point to genes for which literature studies show that a different gene is strongly incriminated in the disease, most likely rendering the prediction wrong (annotated with ”1#” in Supplementary Table 2 online). Of these, 11 are the best-scoring candidate in the interval and 5 score second or lower. When considering only the best-scoring candidate in each interval (as we have done in the benchmark), 25 are most likely true positives and 11 are most likely negatives. Thus, the precision is 69% - even better than the precision in the benchmark, where predictions above 0.2 have a precision of 49%. Sixty-six of the candidates belong to intervals where there is no evidence in the literature regarding a gene(s) that contributes to the pathology. We consider these as novel candidates. All complexes underlying the candidates scoring 0.1 or above are available for download from the database supporting this work.

To exemplify the candidate protein complexes underlying the scoring of the Bayesian predictor, we present four case studies of the novel candidates from Supplementary Table 2 online. Similar analysis can be carried out for all 506 complexes in the data set, pointing to specific approaches toward validation of the proposed relationships.

Case studies

Retinitis pigmentosa is a clinically and genetically heterogeneous group of disorders. Common traits are night blindness, constricted visual field and retinal dystrophy. In an associated interval on 2p15-p11 (ref. [105]), the Bayesian predictor points to LOC130951 with a score of 0.5232. This protein is uncharacterized but evolutionarily conserved, and it is putatively involved in the disease based on an interaction with CRX25, [127] (Fig. 3a). CRX is a homeobox transcription factor known to be involved in retinitis pigmentosa and cone rod dystrophy [243]. The candidature of LOC130951 is not obvious, and because both interaction studies reporting the interaction to CRX are large scale, including thousands of interactions, it seems unlikely that LOC130951 would have been chosen as a suitable candidate by manual investigation of the interval.

Epithelial ovarian cancer arises as a result of genetic alterations in the ovarian surface epithelium. In an associated interval on 3p25-p22 (ref. [235]), the Bayesian predictor points to Fanconi anemia group D2 protein (FANCD2) with a score of 0.9981. This protein is placed in a complex with breast cancer type 2 susceptibility protein (BRCA2), breast cancer type 1 susceptibility protein (BRCA1) and nibrin isoform 1 (NBN), all of which are involved in ovarian cancer, breast cancer or chromosomal instability disorders [66, 184, 46, 161] (Fig. 3b). Furthermore, other proteins involved in cancer can be identified in the complex (Supplementary Data and Supplementary Fig. 4 online). FANCD2 is part of the BRCA pathway in cisplatin-sensitive cells [161] and is known to be involved in different types of cancer [259]. However, to our knowledge, a mutation in this gene has never been demonstrated in epithelial ovarian cancer, and we consider it to be a likely candidate in epithelial ovarian cancer in families with linkage to 3p22-p25.

Inflammatory bowel disease is characterized by chronic, relapsing intestinal inflammation. In an associated interval on 6p [65, 107], the Bayesian predictor points to receptor-interacting serine/threonine protein kinase (RIPK1) as the most likely candidate with a score of 0.9984 (Fig. 3c). The candidate complex includes the signaling proteins tumor necrosis factor receptor 2 (TNFRSF1B), tumor necrosis factor precursor (TNF) and tumor necrosis factor receptor precursor (TNFRSF1A), all known to be associated with inflammatory bowel disease or other inflammatory disorders. Furthermore, other proteins involved in inflammation and immune responses can be observed in the complex (Supplementary Data and Supplementary Fig. 5 online). We thus identified a positional candidate, which is placed centrally in a complex of proteins known to be involved in inflammatory bowel disease and other types of inflammation. We note that RIPK1 lies 20.6 Mb from the closest high-resolution marker published. However, considering that all of 6q was screened for candidates, and that several genes lying far from the published markers are most likely true predictions in Supplementary Table 2 online, we believe that RIPK1 is a very likely candidate involved in inflammatory bowel disease.

Amyotrophic lateral sclerosis (ALS) with frontotemporal dementia is a degenerative motor neuron disorder characterized by muscular atrophy, progressive motor neuron function loss and bulbar paralysis. In many families, hereditary ALS is associated with frontotemporal dementia and linkage has been shown to an area on 9q21-q22 (ref. [122]). Here, the Bayesian predictor points to two likely candidates: bicaudal D homolog 2 (BICD2) and cytoplasmic isoleucyl-tRNA synthetase (IARS), scoring 0.4351 and 0.2154, respectively. Although BICD2 is scored highest, both candidates are awarded good scores and are plausible candidates for contributing to ALS associated with dementia. However, investigation of the candidate complexes suggests that BICD2 is more likely to be involved in non-familial ALS not associated with dementia, because it is part of a complex with dynactin, which is associated with ALS without dementia. IARS is in a complex with superoxide dismutase 1, a protein known to be involved in familial ALS47 including dementia (Fig. 3d). Also, the IARS complex contains molecular chaperones and other proteins that have been connected to the disease and other types of dementias (Supplementary Fig. 6 online), and the interaction data underlying the complex is highly reproducible (Supplementary Data online). Both candidates are likely, but the candidate complex underlying IARS is seemingly more relevant to familial ALS, and it is plausible that IARS could be involved in the disease in families with linkage to 9q21-q22. Because little is known about this disorder, the complex revealed here is an interesting new lead concerning its underlying causes.

These case studies indicate the value of data mining our phenome-interactome network and integrating interaction data across multiple organisms for positional candidate prioritization. In the case of retinitis pigmentosa and ALS with frontotemporal dementia, the predictor identifies non obvious candidates in novel putative complexes supported by a network of reproducible interaction data from humans and multiple model organisms. In the cases of inflammatory bowel disease and epithelial ovarian cancer, we identify partly characterized complexes, where several members are known to be involved in the patient phenotype. However, because there are ˜500 positional candidates in the case of inflammatory bowel disease, it would require extensive literature studies to reveal this network and candidate by manual data integration. We thus believe that RIPK1 would probably not have been identified as a good candidate despite prior knowledge of its involvement in a known network contributing to inflammatory responses.

DISCUSSION

We have recently witnessed the emergence of integrative methods for identifying probable disease genes in linkage intervals associated with disease based on data integration involving, for example, Gene Ontology categories and expression data [2, 91, 92, 262, 210, 211, 182, 269, 270, 123, 93]. Traditionally these methods are compared by measuring average fold enrichment of positional probability (Supplementary Methods online). If a method ranks the true candidate in the top 10% of all candidates in 50% of the linkage intervals, there is a tenfold enrichment in the successful predictions intervals and fivefold enrichment on average. We show that our method increases the probability 108.8 times for the successful predictions and 23.1 times on average, significantly outperforming the other computational methods for positional candidate prioritization, which report 5.6-31.2 times enrichment in the successful linkage intervals to 3.8 to 19.4 times enrichment on average (Supplementary Table 3 online). The most common failure of our method to correctly identify the disease gene results from the inability to find interaction partners associated with a similar phenotype as the relevant protein. This could result from either a lack of data or the failure of these proteins to interact with proteins involved in similar phenotypes. In 75% of these cases, failure to identify another candidate scoring over 0.1 eliminates the possibility of an incorrect prediction.

Our ability to assign candidates to high-confidence protein complexes and rank these complexes in terms of phenomics has permitted us to present a first draft of 506 protein complexes associated with human disease. The success of our method can be attributed to a combination of factors. First, we integrate experimental protein interaction data with a phenotype similarity scheme, thereby taking advantage of the complete clinical spectrum of related human diseases. Also, we use high-confidence protein complexes for identifying novel candidates, thus ensuring that we take advantage of the full protein network context of the candidate, which we show is well suited for functional association of proteins with diseases. Only three of the previously published methods use protein interaction data [91, 203, 3]. Whereas one [203] relies completely on unscored binary interaction pairs to identify candidates in identical diseases, others [91, 3] incorporate unscored human protein interaction data as one of the weaker sources of information. The two latter methods do not take advantage of cross-species integration of interaction data and none of the three integrate phenotypic descriptions as we have done. Furthermore, two approaches [203, 91] search only for candidates implicated in identical diseases and do not take advantage of information from different diseases with a phenotypic overlap. Another method22 relies on provision of a training set and could theoretically be trained using proteins involved in non identical but overlapping phenotypes. These methods report 10.0-15.4 times enrichment in the successful linkage intervals and 5.0-10.0 times enrichment on average (Supplementary Table 3 online). All three methods are innovative and of high quality, but the difference in performance can readily be explained by recalling that the use of high-confidence protein complexes and data about overlapping phenotypes is much better at inferring functional associations than the search for unscored single-interaction partners involved in identical phenotypes only. The complexes generated in the training and validation of the method provide a valuable resource for further investigations by researchers investigating these diseases, because the complexes place the disease-causing proteins in a functional context relative to other disease-associated proteins. We have created a database of these two data sets (available from `http://www.cbs.dtu.dk/suppl/dgf/`) providing a draft of 506 putative human disease complexes, determined by the current resolution of data. Our validation shows that the score associated with each complex can be used as a reliable indication of the quality of the data underlying the complex.

METHODS

Design choices of the Bayesian predictor.

We have strived to make optimal design choices to guarantee the quality of the methodology. First, for the phenotype similarity score, we opted for the UMLS vocabulary, because it is a well-known resource for this type of analysis, and MMTx for the term mapping. There are some limitations when using MMTx and UMLS (see Supplementary Methods online), but we concluded that these are well suited for our analysis, and improvement of these resources is beyond the scope of this work. Second, we chose term frequency-inverse document frequency (tf-idf) as the term-weighting strategy. Compared with unweighted vectors and idf term weighting, tf-idf performed better (Supplementary Fig. 7 online). Third, we used the cosine similarity measure between phenotype vectors, because it is a well-accepted similarity measure for weighted-term vectors. We demonstrate the robustness of this measure on phenotype vectors constructed from a different text source, weighting method and vocabulary (Supplementary Fig. 8 online). Finally, for reporting likely candidates, a threshold of 0.1 on the Bayesian score was chosen on the basis of our benchmark. Using these design choices we created a Bayesian model that was trained and validated using five-fold cross-validation. Additionally, the model was thoroughly optimized to get the optimal separation of signal to noise from the phenotype similarity scheme, the protein interaction data and the other parameters in the model. This was done using a genetic algorithm (Supplementary Methods online).

Filtering irrelevant semantic types from UMLS.

The UMLS vocabulary was manually checked for semantic types that were obviously not clinically relevant (for example, STY|T066|Machine Activity, STY|T068|Human-caused Phenomenon or Process, STY|T093|Health Care Related Organization, STY|T097|Professional or Occupational Group). Terms belonging to these semantic types were filtered out and do not appear in the phenotype vectors. This procedure helps in limiting the phenotype vectors to relevant medical terms to as large an extent as possible.

Phenotype similarity scores.

Both the text and clinical synopsis parts of each OMIM record were parsed with MMTx (`http://mmtx.nlm.nih.gov/`) (for a discussion on the recall, precision and well documented problems of MMTx see Supplementary Methods online) to find the occurrence of medical terms in a subset of the UMLS vocabulary [33], where a number of obviously non clinical semantic type categories had been removed. Phenotype vectors for each record were constructed so that the value of each dimension in the vector represents the number of occurrences of that term in that particular record. Because many relevant terms (for example, mental retardation) are very frequent in OMIM, we also assigned a weight to every extracted term in a phenotype vector. This was done by comparing the frequency with which the term was used in the record in question to its mention the term in all records (that is, all of OMIM). This weight is called tf-idf [212] (Supplementary Methods online) and markedly improves the predictive quality of the data (Supplementary Methods online). Furthermore, this procedure normalizes the term weight using the length of the specific record and the total length of all records. This normalization reduces negative bias in relation to short records, and positive bias in relation to long records. Once vectors for all records had been constructed, pairwise similarity was calculated as the cosine of the angle between the OMIM vectors after normalization [229]. We used the cosine measure as a natural similarity score for two vectors, because it is a standard measure used in this type of text-mining analysis and it is fast to calculate. We note a small bias against some of the phenotype vectors used to predict because of less well curated and described phenotype records in the prediction set than in the benchmarking set (Supplementary Table 4 online). We believe this bias is largely caused by less extensive annotation by the OMIM curators of records describing loci where the disease gene has not been identified. The result is fewer predictions than expected from the benchmark. However, it is important to note that the predictions we do get are of equal quality to the benchmarking case, because the posterior probability score relies on the quality of the data used for the prediction.

Validating the phenotype similarity score.

To investigate to what extent our phenotype vector cosine scores could correctly assign phenotype similarity between scored records, we fitted a curve of the score against the overlap in OMIM record pairs that had a high degree of phenotypic overlap (Supplementary Methods online). The curve shows that the computational phenotype similarity score is directly correlated to the probability of overlap with these record pairs (Supplementary Fig. 2 online)

Constructing a scored human protein interaction network.

Protein interaction data were downloaded from MINT [296], BIND [11], IntAct [117], KEGG annotated protein-protein interactions (PPrel), KEGG Enzymes involved in neighboring steps (ECrel) [144] and Reactome proteins involved in the same complex, indirect complex, reaction or neighboring reaction [141]. All human data were pooled, and to increase the coverage of interactions, interolog data (the transfer of protein interactions between orthologous protein pairs in different organisms) [278] were included by a method similar to that reported by Lehner and Fraser [168]. Interactions were transferred from 17 eukaryotic organisms and added to the network. Orthology was assigned using the Inparanoid database [199] with strict thresholds. To obtain a global interaction score for all interactions in the network, we constructed a probabilistic protein interaction score that took into account the topology of the interaction network surrounding the interaction, the experimental setup (large-scale vs. small-scale) and the number of different publications in which the interaction had been detected (Supplementary Methods online).

Making a virtual pull-down.

A virtual pull-down of a given protein was done by querying the interaction network for all interactions of the protein (and subsequently all interactions between the interacting proteins) and only retaining the interactions over a given score threshold as defined by the genetic algorithm in the training steps of the Bayesian predictor. This means that the resulting interactions all are of high confidence and supported by network topology, different publications, reliable small-scale interaction experiments, reproducibility or a combination of these.

Identifying proteins involved in diseases in the candidate complexes.

Ensembl Mart(`http://dec2005.archive.ensembl.org/Multi/martview`) was used to associate proteins to phenotypes (MIMS) and identify proteins involved in disease in the candidate complexes.

Making the benchmarking cases.

A list of 3,256 disease genes was initially downloaded from the Disease Gene table in GeneCards (`http://nciarray.nci.nih.gov/cards/`). GeneCards mines several different databases, including OMIM, for text describing the disease genes in this table. For some of the disease genes the entries in GeneCards are sentences, originating from OMIM, specifically stating that defects in particular genes lead to particular diseases. To exclude genes associated to diseases by circumstantial evidence, and only include genes in which genetic defects were known to be causative in relation to the particular disorders, we included genes in the benchmarking set only if GeneCards had found such sentences in OMIM in relation to the gene. Because OMIM is a database manually curated by disease experts, we consider such statements from OMIM to be trustworthy. However, to double-check that no mistakes were made by GeneCards in the extraction procedure, or in the curation process by OMIM, we randomly selected 50 of these statements and manually checked (i) that such statements were actually present in the relevant OMIM files and (ii) that the statements were supported by cited literature. In these 50 cases no discrepancies were found, and this investigation led us to consider that all of the statements are correct. This procedure led to a subset of 963 genes and their corresponding proteins. These genes and proteins were associated with their respective phenotypes using GeneCards references to OMIM diseases. This showed that the 963 genes are involved in 1,404 distinct phenotypes, which were used for the training and validation of the Bayesian predictor. Benchmarking cases were made by associating the genes to distinct phenotypes using the annotation in GeneCards and by assigning the genes to artificial linkage intervals. This was done by including a random number of genes upstream and downstream of the known disease gene. The interval sizes were randomized so that they have a distribution similar to the intervals in OMIM morbidmap, for which no gene has been identified, leading to an average of 108.8 genes in each of the 1,404 linkage intervals.

Training and validating the Bayesian model.

Training and benchmarking of the Bayesian model were done by fivefold cross-validation on the benchmarking set. The set of 1,404 benchmarking cases was split into five sets and the Bayesian model trained and optimized on four of these fractions (Supplementary Methods online). Subsequently, the optimized model was used to rank candidates in benchmarking cases made on the last fifth of the data set. This was done for all combinations of the five fractions. The benchmarking results can be seen in Supplementary Table 5 online.

Bayesian disease gene predictor.

The goal is to compute, for each candidate in a critical interval, the probability that this is the disease-related protein. High probabilities should be assigned to candidates that interact with one or more proteins involved in disorders that are phenotypically similar to the one being investigated. This logic is expressed in the form of a probabilistic model and we use Bayes' theorem to compute the probabilities. The model includes parameters for (1) the probability that a candidate protein has any reported interaction partners, (2) protein interaction score, (3) the number of interaction partners that are involved in similar disorders, and (4) computational phenotype similarity score. All parameters are estimated from our data sets for both disease- and non-disease-associated genes, where we see that the parameter values are different in the two cases. The probability, that protein number i (among N candidates) is the disease associated one, is computed as follows:

$$P(dis = i|DATA) = \frac{P(DATA|dis = i) \times P(dis = i)}{\sum_{j=1}^{N} P(DATA|dis = j) \times P(dis = j)}$$

Where $P(dis = i|DATA)$ is the *posterior* probability that candidate number i is the disease-related protein after evaluating all the data. $P(dis = i)$ is the *prior* probability that candidate number i is the disease causing protein, before evaluating any data. The prior value was set to $\frac{1}{N}$ for all candidates. The term $P(DATA|dis = i)$ is the probability of value was set to obtaining the observed data if candidate number i was in fact the correct one. This likelihood is computed from the interaction data and any associated phenotype descriptions, and using the estimated parameters, in a straightforward manner (Supplementary Materials on-line).

Case studies.

Case studies were made by downloading complex data available for all putative disease complexes (http://www.cbs.dtu.dk/suppl/dgf/) and creating an interactive graph in the free software cytoscape (http://www.cytoscape.org/). Data in these files combined with literature studies were used to generate the hypotheses. More data on the case studies can be found in (Supplementary Material online). Proteins are named by using the corresponding gene name according to HUGO gene nomenclature http://www.gene.ucl.ac.uk/nomenclature/.

ACKNOWLEDGMENTS

The authors wish to thank Ulrik de Lichtenberg and Thomas Skøt Jensen for critical reading of the manuscript, editing and for help in developing the protein interaction score and Daniel Hvidberg Aaen Hansen for assisting in the analysis of the protein complexes. Z.M.S. is supported by an EU Biosapiens (NoE), FP6 grant. Wilhelm Johannsen Centre for Functional Genome Research is established by the Danish National Research Foundation.

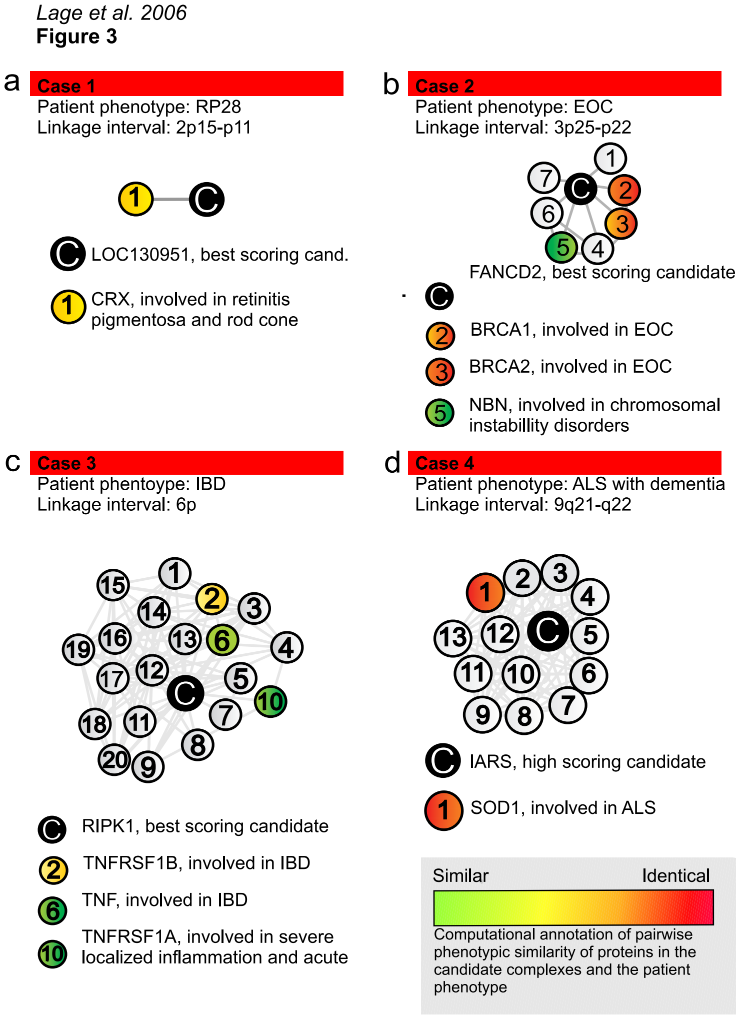

Lage et al. 2006
Figure 3
a
Case 1
Patient phenotype: RP28
Linkage interval: 2p15-p11
1
C
C LOC130951, best scoring cand.
1 CRX, involved in retinitis pigmentosa and rod cone
b
Case 2
Patient phenotype: EOC
Linkage interval: 3p25-p22
1
7
C
2
6
3
5
4
FANCD2, best scoring candidate
C
2 BRCA1, involved in EOC
3 BRCA2, involved in EOC
5 NBN, involved in chromosomal instabillity disorders
c
Case 3
Patient phentoype: IBD
Linkage interval: 6p
15
1
2
14
3
19
16
13
6
4
17
12
5
C
10
7
18
11
8
20
9
C RIPK1, best scoring candidate
2 TNFRSF1B, involved in IBD
6 TNF, involved in IBD
10 TNFRSF1A, involved in severe localized inflammation and acute
d
Case 4
Patient phenotype: ALS with dementia
Linkage interval: 9q21-q22
1
2
3
4
13
12
C
5
11
10
6
9
8
7
C IARS, high scoring candidate
1 SOD1, involved in ALS
Similar
Identical
Computational annotation of pairwise phenotypic similarity of proteins in the candidate complexes and the patient phenotype

Figure 3: Case studies of four candidate complexes. These candidate complexes are virtually pulled-down with the best scoring candidate in retinitis pigmentosa 28 (RP28) (a), epithelial ovarian cancer (EOC) (b), inflammatory bowel disease (IBD) (c) and a high scoring candidate in amyotrophic lateral sclerosis (ALS) with frontotemporal dementia (d), respectively. The black proteins (C) are the high scoring candidates in the four disorders, numbered proteins are proteins interacting with the candidate proteins. Colored nodes are proteins identified by our phenotype association scheme, and grey proteins are not known to be involved in the disorder.

Chapter 5

Epilogue

Reading the complete thesis, it may sound as if the 3 years I spent at CBS working on these projects was a completely planned and coordinated effort. In reality that is far from the truth, as I suppose is true for most research.

My main focus through the duration of my Ph.D. was always protein-protein interaction prediction. The work on Internet technologies and data sharing standards in chapter 2 is a result of CBS responsibilities as a part of BioSapiens NoE. Implementing the distributed annotation system (DAS) in various proteomics labs around Europe opened the possibility of integrated functional research of the ENCODE complement (section 2.1.3), where analysis of splice variants yielded surprising results about the functional and structural diversity of gene products from individual loci. Computational analysis from a dozen labs around Europe were performed and integrated in a matter of days, yet again proving the importance of being on line today.

When our attempts at protein-protein interaction (PPI) prediction did not bear anticipated fruits, we looked more into the practical use of PPI data, looking globally at the resulting interactome instead of individual interactions. Especially we focused on chromosomal interaction maps to see if we could find links between interaction networks and the ability to survive trisomy. I became a part of a disease gene finding project, which focused on PPI data. In that framework we successfully attached phenotype annotations to the network of interactions to identify genes with disease phenotypes.

Diseases with a simple Mendelian inheritance model, such as cystic fibrosis and Huntington's disease are rare. In the case of complex diseases there are several genes/proteins whose specific alleles incrementally add to the risk of developing the disease. Diabetes is one such disease where a suspected 5-10 (or even more) genes are involved [57]. Despite ever higher genotype resolutions, ever faster sequencing techniques, linking disease phenotypes and specific genes continues to be a tough task, one where systems biology may provide valuable understanding by identifying complex relationships between the genes implicated in a particular disease [153]. One field where there is room for considerable improvements is phenomics. The success of the simple phenotype vector assignment-and-overlap framework, used to identify related genetic diseases in section 4.3 shows the possibilities for use and integration of such data in biology. The disease gene finding effort proves that the amount of data being generated by high throughput experiments has reached critical mass for systems biology and that the resulting networks are now saturated enough for practical use. The ultimate target is a future where we, not just understand the complex systems within the cell, but rather are able to manipulate these with precision bioengineering.

High throughput experiments are not only allowing us to analyse data in a systems fashion. Being able to integrate data in such way helps us to identify the pieces that stand out and to go back and scrutinise the those pieces. The genome sequence has already been uncovered, and as 99.9% of it is identical in us all, we have to look elsewhere for the cause of variation in phenotype and disease. Recent studies indicate that quantitative experiments of expression may hold the key to our phenotypes, both with regard to regulation [245] and copy number variation in DNA [218]. Analyses of primary sequence are not yet yielding answers to all the *whys*, but rather pointing to new paths to explore. Therefore we hope for systems biology to bridge the gap between genotype and phenotype. There is a lot of work to be done but we are certainly making progress.

Bibliography

[1] The international hapmap project. *Nature*, 426:789–796, Dec 2003.

[2] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6:55, Mar 2005.

[3] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24:537–544, May 2006.

[4] R. Albert, H. Jeong, and A. L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:378–382, Jul 2000.

[5] P. Aloy and R. B. Russell. The third dimension for protein interactions and complexes. *Trends Biochem Sci*, 27:633–638, Dec 2002.

[6] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402, Sep 1997.

[7] L. Alvesalo and A. de la Chapelle. Permanent tooth sizes in 46,xx-males. *Ann Hum Genet*, 43:97–102, Oct 1979.

[8] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proc AMIA Symp*, pages 17–21, 2001.

[9] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone. Comment on “network motifs: simple building blocks of complex networks” and “su-

perfamilies of evolved and designed networks". *Science*, 305:1107; author reply 1107, Aug 2004.

[10] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25:25–29, May 2000.

[11] G. D. Bader, D. Betel, and C. W. V. Hogue. Bind: the biomolecular interaction network database. *Nucleic Acids Res*, 31:248–250, Jan 2003.

[12] G. D. Bader and C. W. V. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 20:991–997, Oct 2002.

[13] G. D. Bader and C. W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, Jan 2003.

[14] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, 22(1):78–85, Jan 2004.

[15] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 33 Database Issue:D154–9, Jan 1 2005.

[16] P. Baldi and S. Brunak. *Bioinformatics.* MIT Press, 1998.

[17] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–24, May 2000.

[18] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5:101–113, Feb 2004.

[19] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–12, Oct 15 1999.

[20] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The pfam protein families database. *Nucleic Acids Res*, 32:D138–D141, Jan 2004.

[21] B. J. Baty, B. L. Blackburn, and J. C. Carey. Natural history of trisomy 18 and trisomy 13: I. growth, physical assessment, medical histories, survival, and recurrence risk. *Am J Med Genet*, 49:175–188, Jan 1994.

[22] A. Ben-Hur and W. S. Noble. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7 Suppl 1:S2, Mar 2006.

[23] J. D. Bendtsen, L. J. Jensen, N. Blom, G. von Heijne, and S. Brunak. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel*, 17(4):349–56, Apr 2004.

[24] J. D. Bendtsen, H. Nielsen, G. V. Heijne, and S. Brunak. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, 340(4):783–95, Jul 16 2004.

[25] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. Genbank. *Nucleic Acids Res*, 33(Database issue):D34–8, Jan 1 2005.

[26] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28:235–242, Jan 2000.

[27] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American Magazine*, May 01 2001.

[28] D. Betel, R. Isserlin, and C. V. W. Hogue. Analysis of domain correlations in yeast protein complexes. *Bioinformatics*, 20 Suppl 1:I55–I62, Aug 4 2004.

[29] E. Birney, D. Andrews, M. Caccamo, Y. Chen, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, S. Graf, M. Hammond, J. Herrero, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, F. Kokocinski, E. Kulesha, D. London, I. Longden, C. Melsopp, P. Meidl, B. Overduin, A. Parker, G. Proctor, A. Prlic, M. Rae, D. Rios, S. Redmond, M. Schuster, I. Sealy, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith,

A. Stabenau, J. Stalker, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and T. P. J. Hubbard. Ensembl 2006. *Nucleic Acids Res*, 34(Database issue):D556–61, Jan 1 2006.

[30] D. L. Black. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, 103:367–370, Oct 2000.

[31] N. Blom, S. Gammeltoft, and S. Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*, 294(5):1351–62, Dec 17 1999.

[32] J. R. Bock and D. A. Gough. Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–60, May 2001.

[33] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res*, 32:D267–D270, Jan 2004.

[34] S. Boue, I. Letunic, and P. Bork. Alternative splicing and evolution. *Bioessays*, 25:1031–1034, Nov 2003.

[35] T. Boveri. *Zur Frage der Entstehung maligner Tumoren.* Gustav Fisher Verlag, Jena, Germany, 1914.

[36] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nat Genet*, 29:365–371, Dec 2001.

[37] D. Brett, H. Pospisil, J. Valcárcel, J. Reich, and P. Bork. Alternative splicing and genome complexity. *Nat Genet*, 30:29–30, Jan 2002.

[38] B. M. N. Brinkman. Splice variants as cancer biomarkers. *Clin Biochem*, 37:584–594, Jul 2004.

[39] H. G. Brunner and M. A. van Driel. From syndrome families to functional genomics. *Nat Rev Genet*, 5:545–551, Jul 2004.

[40] A. J. Butte and I. S. Kohane. Creation and implications of a phenome-genome network. *Nat Biotechnol*, 24:55–62, Jan 2006.

[41] Y.-D. Cai and K.-C. Chou. Nearest neighbour algorithm for predicting protein subcellular location b y combining functional domain composi-

tion and pseudo-amino acid composition. *Biochem Biophys Res Commun*, 305(2):407–11, May 30 2003.

[42] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res*, 32:D262–D266, Jan 2004.

[43] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. M. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–635, Jun 2006.

[44] S. Carrere and J. Gouzy. Remora: a pilot in the ocean of biomoby web-services. *Bioinformatics*, 22:900–901, Apr 2006.

[45] G. Casari, C. Sander, and A. Valencia. A method to predict functional residues in proteins. *Nat Struct Biol*, 2(2):171–8, Feb 1995.

[46] L. H. Castilla, F. J. Couch, M. R. Erdos, K. F. Hoskins, K. Calzone, J. E. Garber, J. Boyd, M. B. Lubin, M. L. Deshano, and L. C. Brody. Mutations in the brca1 gene in families with early-onset breast and ovarian cancer. *Nat Genet*, 8:387–391, Dec 1994.

[47] J. Cedano, P. Aloy, J. A. Perez-Pons, and E. Querol. Relation between amino acid composition and cellular location of proteins. *J Mol Biol*, 266(3):594–600, Feb 28 1997.

[48] N. Chen, T. W. Harris, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, P. Canaran, J. Chan, C.-K. Chen, W. J. Chen, F. Cunningham, P. Davis, E. Kenny, R. Kishore, D. Lawson, R. Lee, H.-M. Muller, C. Nakamura, S. Pai, P. Ozersky, A. Petcherski, A. Rogers, A. Sabo, E. M. Schwarz, V. K. Auken, Q. Wang, R. Durbin, J. Spieth, P. W. Sternberg, and L. D. Stein. WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res*, 33 Database Issue:D383–9, Jan 1 2005.

[49] K.-C. Chou and Y.-D. Cai. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J Cell Biochem*, 90(6):1250–60, Dec 15 2003.

[50] K. C. Chou and Y. D. Cai. Prediction of protein subcellular locations by-fund-pseaa predictor. *Biochem Biophys Res Commun*, 320(4):1236–9., Aug 6 2004.

[51] H. H. Chun, S. Castellví-Bel, Z. Wang, R. A. Nagourney, S. Plaeger, S. G. Becker-Catania, F. Naeim, R. S. Sparkes, and R. A. Gatti. Tcl-1, mtcp-1 and tml-1 gene expression profile in non-leukemic clonal proliferations associated with ataxia-telangiectasia. *Int J Cancer*, 97:726–731, Feb 2002.

[52] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proc Natl Acad Sci U S A*, 99:15879–15882, Dec 2002.

[53] T. Clark, S. Martin, and T. Liefeld. Globally distributed object identification for biological knowledgebases. *Brief Bioinform*, 5(1):59–70, Mar 2004.

[54] A. M. Cohen and W. R. Hersh. A survey of current work in biomedical text mining. *Brief Bioinform*, 6:57–71, Mar 2005.

[55] B. A. Cohen, R. D. Mitra, J. D. Hughes, and G. M. Church. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet*, 26:183–186, Oct 2000.

[56] R. Cohen and S. Havlin. Scale-free networks are ultrasmall. *Phys Rev Lett*, 90:058701, Feb 2003.

[57] F. S. Collins and V. A. McKusick. Implications of the human genome project for medical science. *JAMA*, 285:540–544, Feb 2001.

[58] T. E. consortium. The encode (encyclopedia of dna elements) project. *Science*, 306:636–640, Oct 2004.

[59] R. G. Côté, P. Jones, R. Apweiler, and H. Hermjakob. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7:97, Feb 2006.

[60] M. O. Dayhoff, R. M. Schwartz, H. R. Chen, L. T. Hunt, W. C. Barker, and B. C. Orcutt. Nucleic acid sequence bank. *Science*, 209(4462):1182, Sep 12 1980.

[61] U. de Lichtenberg, L. J. Jensen, S. Brunak, and P. Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307:724–727, Feb 2005.

[62] U. N. de Lichtenberg. *Bioinformatics and Systems Biology of the Cell Cycle*. PhD thesis, Technical University of Denmark, DTU, 2005.

[63] D. de Solla Price. A general theory of bibliometrics and other cumulative advantage processes. *J Am Soc Inform Sci*, 27:292–306, 1976.

[64] D. J. de Solla Price. Networks of scientific papers. *Science*, 149:510–515, 1965.

[65] B. Dechairo, C. Dimon, D. van Heel, I. Mackay, M. Edwards, P. Scambler, D. Jewell, L. Cardon, N. Lench, and A. Carey. Replication and extension studies of inflammatory bowel disease susceptibility regions confirm linkage to chromosome 6p (ibd3). *Eur J Hum Genet*, 9:627–633, Aug 2001.

[66] I. Demuth, P.-O. Frappart, G. Hildebrand, A. Melchers, S. Lobitz, L. Stöckl, R. Varon, Z. Herceg, K. Sperling, Z.-Q. Wang, and M. Digweed. An inducible null mutant murine model of nijmegen breakage syndrome proves the essential function of nbs1 in chromosomal stability and cell viability. *Hum Mol Genet*, 13:2385–2397, Oct 2004.

[67] M. Deng, S. Mehta, F. Sun, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 12(10):1540–8, Oct 2002.

[68] N. Deshpande, K. J. Addess, W. F. Bluhm, J. C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R. K. Green, J. L. Flippen-Anderson, J. Westbrook, H. M. Berman, and P. E. Bourne. The RCSB protein data bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res*, 33 Database Issue:D233–7, Jan 1 2005.

[69] R. Diestel. *Graph Theory*. Springer-Verlag, Heidelberg, 2005.

[70] G. Divita, T. Tse, and L. Roth. Failure analysis of metamap transfer (mmtx). *Medinfo*, 11:763–767, 2004.

[71] S. Dohkan, A. Koike, and T. Takagi. Prediction of protein-protein interactions using support vector machines. *Proceedings. Fourth IEEE Symposium on Bioinformatics and Bioengineering*, pages 576–83, 2004.

[72] R. D. Dowell, R. M. Jokerst, A. Day, S. R. Eddy, and L. Stein. The distributed annotation system. *BMC Bioinformatics*, 2(1):7, 2001.

[73] R. A. Drysdale, M. A. Crosby, W. Gelbart, K. Campbell, D. Emmert, B. Matthews, S. Russo, A. Schroeder, F. Smutniak, P. Zhang, P. Zhou, M. Zytkovicz, M. Ashburner, de A. Grey, R. Foulger, G. Millburn, D. Sutherland, C. Yamada, T. Kaufman, K. Matthews, A. DeAngelo, R. K. Cook, D. Gilbert, J. Goodman, G. Grumbling, H. Sheth, V. Strelets, G. Rubin, M. Gibson, N. Harris, S. Lewis, S. Misra, and S. Q. Shu. FlyBase: genes and gene models. *Nucleic Acids Res*, 33 Database Issue:D390–5, Jan 1 2005.

[74] P. Duckert, S. Brunak, and N. Blom. Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel*, 17(1):107–12, Jan 2004.

[75] P. Duesberg, R. Li, A. Fabarius, and R. Hehlmann. The chromosomal basis of cancer. *Cell Oncol*, 27:293–318, 2005.

[76] P. Duesberg, R. Li, A. Fabarius, and R. Hehlmann. Aneuploidy and cancer: from correlation to causation. *Contrib Microbiol*, 13:16–44, 2006.

[77] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet*, 18:529–536, Oct 2002.

[78] D. Ekman, S. Light, A. K. Björklund, and A. Elofsson. What properties characterize the hub proteins of the protein-protein interaction network of saccharomyces cerevisiae? *Genome Biol*, 7:R45, 2006.

[79] J. A. Ellis, M. Stebbing, and S. B. Harrap. Significant population variation in adult male height associated with the y chromosome and the aromatase gene. *J Clin Endocrinol Metab*, 86:4147–4150, Sep 2001.

[80] O. Emanuelsson, H. Nielsen, S. Brunak, and von G. Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*, 300(4):1005–16, Jul 21 2000.

[81] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, Nov 4 1999.

[82] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30:1575–1584, Apr 2002.

[83] C. J. Epstein. *Down syndrome, trisomy 21.In: Scriver, C. R.; Beaudet, A. L.; Sly, W. S.; Valle, D. : Metabolic Basis of Inherited Disease.* New York: McGraw-Hill, 1989.

[84] P. Erdös and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

[85] G. Ermak, G. Gerasimov, K. Troshina, T. Jennings, L. Robinson, J. S. Ross, and J. Figge. Deregulated alternative splicing of cd44 messenger rna transcripts in neoplastic and nonneoplastic lesions of the human thyroid. *Cancer Res*, 55:4594–4598, Oct 1995.

[86] T. Etzold and P. Argos. Srs–an indexing and retrieval tool for flat file data libraries. *Comput Appl Biosci*, 9:49–57, Feb 1993.

[87] P. Fariselli, F. Pazos, A. Valencia, and R. Casadio. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem*, 269(5):1356–61, Mar 2002.

[88] R. Favaro, R. G. H. Immink, V. Ferioli, B. Bernasconi, M. Byzova, G. C. Angenent, M. Kater, and L. Colombo. Ovule-specific mads-box proteins have conserved protein-protein interactions in monocot and dicot plants. *Mol Genet Genomics*, 268:152–159, Oct 2002.

[89] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340:245–246, Jul 1989.

[90] A. Flanagan, P. Guy, M. Lubkeman, M. Steiner, M. Reeves, and P. Tollman. A revolution in r&d: How genomics and genetics are transforming the biopharmaceutical industry. *Boston Consulting Group*, 2001.

[91] L. Franke, H. v. Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78:1011–1025, Jun 2006.

[92] L. Franke, H. van Bakel, B. Diosdado, M. van Belzen, M. Wapenaar, and C. Wijmenga. Team: a tool for the integration of expression, and linkage and association maps. *Eur J Hum Genet*, 12:633–638, Aug 2004.

[93] J. Freudenberg and P. Propping. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18 Suppl 2:S110–S115, 2002.

[94] D. Frishman, K. Heumann, A. Lesk, and H. W. Mewes. Comprehensive, comprehensible, distributed and intelligent databases: current status. *Bioinformatics*, 14(7):551–61, 1998.

[95] D. Gabellini, G. D'Antona, M. Moggio, A. Prelle, C. Zecca, R. Adami, B. Angeletti, P. Ciscato, M. A. Pellegrino, R. Bottinelli, M. R. Green, and R. Tupler. Facioscapulohumeral muscular dystrophy in mice overexpressing frg1. *Nature*, 439:973–977, Feb 2006.

[96] M. C. Ganoza, M. C. Kiel, and H. Aoki. Evolutionary conservation of reactions in translation. *Microbiol Mol Biol Rev*, 66(3):460–85, table of contents., Sep 2002.

[97] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, A. Edelmann, M.-A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.-M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, Mar 2006.

[98] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, Jan 2002.

[99] H. Ge, Z. Liu, G. M. Church, and M. Vidal. Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nat Genet*, 29:482–486, Dec 2001.

[100] D. Gilbert. Biomolecular interaction network database. *Brief Bioinform*, 6:194–198, Jun 2005.

[101] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of drosophila melanogaster. *Science*, 302:1727–1736, Dec 2003.

[102] C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen. Co-evolution of proteins with their interaction partners. *J Mol Biol*, 299(2):283–93, Jun 2 2000.

[103] S. M. Gomez, W. S. Noble, and A. Rzhetsky. Learning to predict protein-protein interactions from protein sequences. *Bioinformatics*, 19(15):1875–81, Oct 12 2003.

[104] A. Grigoriev. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res*, 31:4157–4161, Jul 2003.

[105] S. Gu, G. Kumaramanickavel, C. R. Srikumari, M. J. Denton, and A. Gal. Autosomal recessive retinitis pigmentosa locus rp28 maps between d2s1337 and d2s286 on chromosome 2p11-p15 in an indian family. *J Med Genet*, 36:705–707, Sep 1999.

[106] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33:D514–D517, Jan 2005.

[107] J. Hampe, S. H. Shaw, R. Saiz, N. Leysens, A. Lantermann, S. Mascheretti, N. J. Lynch, A. J. MacPherson, S. Bridger, S. van Deventer, P. Stokkers, P. Morin, M. M. Mirza, A. Forbes, J. E. Lennard-Jones, C. G. Mathew, M. E. Curran, and S. Schreiber. Linkage of inflammatory bowel disease to human chromosome 6p. *Am J Hum Genet*, 65:1647–1655, Dec 1999.

[108] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and

M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430:88–93, Jul 2004.

[109] J.-D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol*, 23:839–844, Jul 2005.

[110] D. Hansemann. Ueber asymmetrische zelltheilung in epithelkrebsen und deren biologische bedeutung. *Virchows Arch. Pathol. Anat*, 119:299–326, 1890.

[111] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, 32:D258–D261, Jan 2004.

[112] J. Harrow, F. Denoeud, A. Frankish, A. Reymond, C.-K. Chen, J. Chrast, J. Lagarde, J. G. R. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S. E. Antonarakis, and R. Guigo. Gencode: producing a reference annotation for encode. *Genome Biol*, 7 Suppl 1:S4.1–S4.9, Aug 2006.

[113] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, Dec 1999.

[114] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins. Computational studies of gene regulatory networks: in numero molecular biology. *Nat Rev Genet*, 2:268–279, Apr 2001.

[115] X. He and J. Zhang. Why do hubs tend to be essential in protein networks? *PLoS Genet*, 2:e88, Jun 2006.

[116] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, von Christian Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li,

R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. N. G. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, and R. Apweiler. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–83, Feb 2004.

[117] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. Intact: an open source molecular interaction database. *Nucleic Acids Res*, 32:D452–D455, Jan 2004.

[118] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC genome browser database: update 2006. *Nucleic Acids Res*, 34(Database issue):D590–8, Jan 1 2006.

[119] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sørensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–183, Jan 2002.

[120] R. Hoffmann, M. Krallinger, E. Andres, J. Tamames, C. Blaschke, and A. Valencia. Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE*, 2005:pe21, May 2005.

[121] N. S. Holter, A. Maritan, M. Cieplak, N. V. Fedoroff, and J. R. Banavar. Dynamic modeling of gene expression data. *Proc Natl Acad Sci U S A*, 98:1693–1698, Feb 2001.

[122] B. A. Hosler, T. Siddique, P. C. Sapp, W. Sailor, M. C. Huang, A. Hossain, J. R. Daube, M. Nance, C. Fan, J. Kaplan, W. Y. Hung, D. McKenna-Yasek, J. L. Haines, M. A. Pericak-Vance, H. R. Horvitz, and R. H. Brown. Linkage of familial amyotrophic lateral sclerosis with frontotemporal dementia to chromosome 9q21-q22. *JAMA*, 284:1664–1669, Oct 2000.

[123] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey. Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*, 74:289–298, Mar 2005.

[124] T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and E. Birney. Ensembl 2005. *Nucleic Acids Res*, 33 Database Issue:D447–53, Jan 1 2005.

[125] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, 34:W729–W732, Jul 2006.

[126] T. Ideker and D. Lauffenburger. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol*, 21:255–262, Jun 2003.

[127] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98:4569–4574, Apr 2001.

[128] J. Jacob, J. Haspel, N. Kane-Goldsmith, and M. Grumet. L1 mediated homophilic binding and neurite outgrowth are modulated by alternative splicing of exon 2. *J Neurobiol*, 51:177–189, Jun 2002.

[129] J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. E. J. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak. CAPRI: a critical assessment of PRedicted interactions. *Proteins*, 52(1):2–9, Jul 1 2003.

[130] R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 12:37–46, Jan 2002.

[131] R. Jansen, N. Lan, J. Qian, and M. Gerstein. Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics*, 2(2):71–81, 2002.

[132] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–53, Oct 17 2003.

[133] V. Janssens, C. van Hoof, E. Martens, I. de Baere, W. Merlevede, and J. Goris. Identification and characterization of alternative splice products encoded by the human phosphotyrosyl phosphatase activator gene. *Eur J Biochem*, 267:4406–4413, Jul 2000.

[134] L. J. Jensen and P. Bork. Quality analysis and integration of large-scale molecular data sets. *Drug Discovery Today: TARGETS*, 3:51–56, 2004.

[135] L. J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H. H. Staerfeldt, K. Rapacki, C. Workman, C. F. A. Andersen, S. Knudsen, A. Krogh, A. Valencia, and S. Brunak. Prediction of human protein function from post-translational modification s and localization features. *J Mol Biol*, 319(5):1257–65, Jun 21 2002.

[136] L. J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, 7:119–129, Feb 2006.

[137] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, May 2001.

[138] T. Joachims. Training linear svms in linear time. pages 1–10, 2006.

[139] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20., Jan 9 1996.

[140] S. Jones and J. M. Thornton. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol*, 272(1):121–32, Sep 12 1997.

[141] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis,

E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33:D428–D432, Jan 2005.

[142] K. Julenius, A. Molgaard, R. Gupta, and S. Brunak. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology*, 15(2):153–64, Feb 2005.

[143] A. S. Juncker, H. Willenbrock, V. G. Heijne, S. Brunak, H. Nielsen, and A. Krogh. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci*, 12(8):1652–62, Aug 2003.

[144] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res*, 34:D354–D357, Jan 2006.

[145] E. F. Keller. Revisiting "scale-free" networks. *Bioessays*, 27:1060–1068, Oct 2005.

[146] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A*, 100:11394–11399, Sep 2003.

[147] P. Kemmeren, N. L. van Berkum, J. Vilo, T. Bijma, R. Donders, A. Brazma, and F. C. P. Holstege. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell*, 9:1133–1143, May 2002.

[148] R. Khanin and E. Wit. How scale-free are biological networks. *J Comput Biol*, 13:810–818, Apr 2006.

[149] L. Kiemer, J. D. Bendtsen, and N. Blom. Netacet: prediction of N-terminal acetylation sites. *Bioinformatics*, 21(7):1269–70, Apr 1 2005.

[150] W. K. Kim, J. Park, and J. K. Suh. Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. volume 13, pages 42–50, Object Interaction Technologies Inc., Room 201 Jueun Building, 29-4 Jamwon-dong, Seocho-gu, Seoul 137-904, Korea. dimlightoitek.com, 2002.

[151] S. Kirsch, B. Weiss, M. De Rosa, T. Ogata, G. Lombardi, and G. A. Rappold. Fish deletion mapping defines a single location for the y chromosome stature gene, gcy. *J Med Genet*, 37:593–599, Aug 2000.

[152] S. Kishore and S. Stamm. The snorna hbii-52 regulates alternative splicing of the serotonin receptor 2c. *Science*, 311:230–232, Jan 2006.

[153] H. Kitano. Computational systems biology. *Nature*, 420:206–210, Nov 2002.

[154] T. Kodadek. Protein microarrays: prospects and problems. *Chem Biol*, 8:105–115, Feb 2001.

[155] J. O. Korbel, T. Doerks, L. J. Jensen, C. Perez-Iratxeta, S. Kaczanowski, S. D. Hooper, M. A. Andrade, and P. Bork. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol*, 3:e134, May 2005.

[156] A. Kozarova, S. Petrinac, A. Ali, and J. W. Hudson. Array of informatics: Applications in modern research. *J Proteome Res*, 5:1051–1059, May 2006.

[157] S. Kriaucionis and A. Bird. The major form of mecp2 has a novel n-terminus generated by alternative splicing. *Nucleic Acids Res*, 32:1818–1823, Mar 2004.

[158] E. V. Kriventseva, I. Koch, R. Apweiler, M. Vingron, P. Bork, M. S. Gelfand, and S. Sunyaev. Increase of functional diversity by alternative splicing. *Trends Genet*, 19:124–128, Mar 2003.

[159] M. Kurkela, S. Mörsky, J. Hirvonen, R. Kostiainen, and M. Finel. An active and water-soluble truncation mutant of the human udp-glucuronosyltransferase 1a9. *Mol Pharmacol*, 65:826–831, Apr 2004.

[160] T. la Cour, L. Kiemer, A. Molgaard, R. Gupta, K. Skriver, and S. Brunak. Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng Des Sel*, 17(6):527–36, Jun 2004.

[161] J. M. Lancaster, R. Wooster, J. Mangion, C. M. Phelan, C. Cochran, C. Gumbs, S. Seal, R. Barfoot, N. Collins, G. Bignell, S. Patel, R. Hamoudi, C. Larsson, R. W. Wiseman, A. Berchuck, J. D. Iglehart, J. R. Marks, A. Ashworth, M. R. Stratton, and P. A. Futreal. Brca2 mutations in primary breast and ovarian cancers. *Nat Genet*, 13:238–240, Jun 1996.

[162] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda,

W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh,

F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and J. Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, Feb 2001.

[163] M. Lappe and L. Holm. Unraveling protein interaction networks with near-optimal efficiency. *Nat Biotechnol*, 22(1):98–103, Jan 2004.

[164] D. A. Lauffenburger. Cell signaling pathways as control modules: complexity for simplicity? *Proc Natl Acad Sci U S A*, 97:5031–5033, May 2000.

[165] D. J. Law, E. M. Labut, R. D. Adams, and J. L. Merchant. An isoform of zbp-89 predisposes the colon to colitis. *Nucleic Acids Res*, 34:1342–1350, Mar 2006.

[166] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–8, Nov 26 2004.

[167] J. M. Lee and E. L. L. Sonnhammer. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res*, 13:875–882, May 2003.

[168] B. Lehner and A. G. Fraser. A first-draft human protein-interaction map. *Genome Biol*, 5:R63, Aug 2004.

[169] H. Liang and L. F. Landweber. A genome-wide study of dual coding regions in human alternatively spliced genes. *Genome Res*, 16:190–196, Feb 2006.

[170] J. Lim, T. Hao, C. Shaw, A. J. Patel, G. Szabó, J.-F. Rual, C. J. Fisk, N. Li, A. Smolyar, D. E. Hill, A.-L. Barabási, M. Vidal, and H. Y. Zoghbi. A protein-protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell*, 125:801–814, May 2006.

[171] S. L. Lo, C. Z. Cai, Y. Z. Chen, and M. C. Chung. Effect of training datasets on support vector machine prediction of protein-protein interactions. *Proteomics*, 5(4):876–84., Mar 2005.

[172] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14:1675–1680, Dec 1996.

[173] A. J. Lopez. Alternative splicing of pre-mrna: developmental consequences and mechanisms of regulation. *Annu Rev Genet*, 32:279–305, 1998.

[174] G. MacBeath and S. L. Schreiber. Printing proteins as microarrays for high-throughput function determination. *Science*, 289:1760–1763, Sep 2000.

[175] H. Mamitsuka. Essential latent knowledge for protein-protein interactions: Analysis by an unsupervised learning approach. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 2(2):119–130, 2005.

[176] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–3, Jul 30 1999.

[177] S. Marguerat, T. S. Jensen, U. de Lichtenberg, B. T. Wilhelm, L. J. Jensen, and J. Bähler. The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast. *Yeast*, 23:261–277, Mar 2006.

[178] E. Marshall. Bermuda rules: community spirit, with teeth. *Science*, 291:1192, Feb 2001.

[179] S. Martin, D. Roe, and J.-L. Faulon. Predicting protein-protein interactions using signature products. *Bioinformatics*, Aug 19 2005.

[180] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, May 2002.

[181] S. Maslov and K. Sneppen. Computational architecture of the yeast regulatory network. *Phys Biol*, 2:S94–S100, Nov 2005.

[182] M. Masseroli, O. Galati, and F. Pinciroli. Gfinder: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res*, 33:W717–W723, Jul 2005.

[183] K. Matsushita, T. Tomonaga, H. Shimada, A. Shioya, M. Higashi, H. Matsubara, K. Harigaya, F. Nomura, D. Libutti, D. Levens, and T. Ochiai. An essential role of alternative splicing of c-myc suppressor fuse-binding protein-interacting repressor in carcinogenesis. *Cancer Res*, 66:1409–1417, Feb 2006.

[184] S. Matsuura, H. Tauchi, A. Nakamura, N. Kondo, S. Sakamoto, S. Endo, D. Smeets, B. Solder, B. H. Belohradsky, V. M. Der Kaloustian, M. Os-

himura, M. Isomura, Y. Nakamura, and K. Komatsu. Positional cloning of the gene for nijmegen breakage syndrome. *Nat Genet*, 19:179–181, Jun 1998.

[185] B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta*, 405:442–451, Oct 1975.

[186] C. v. Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33 Database Issue:D433–7, Jan 1 2005.

[187] H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, and A. Ruepp. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, 32 Database issue:D41–4, Jan 1 2004.

[188] T. Mijalski, A. Harder, T. Halder, M. Kersten, M. Horsch, T. M. Strom, H. V. Liebscher, F. Lottspeich, M. H. de Angelis, and J. Beckers. Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues. *Proc Natl Acad Sci U S A*, 102:8621–8626, Jun 2005.

[189] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298:824–827, Oct 2002.

[190] G. R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T. M. Raghavan, S. Menon, G. Hanumanthu, M. Gupta, S. Upendran, S. Gupta, M. Mahesh, B. Jacob, P. Mathew, P. Chatterjee, K. S. Arun, S. Sharma, K. N. Chandrika, N. Deshpande, K. Palvankar, R. Raghavnath, R. Krishnakanth, H. Karathia, B. Rekha, R. Nayak, G. Vishnupriya, H. G. M. Kumar, M. Nagini, G. S. S. Kumar, R. Jose, P. Deepthi, S. S. Mohan, T. K. B. Gandhi, H. C. Harsha, K. S. Deshpande, M. Sarker, T. S. K. Prasad, and A. Pandey. Human protein reference database–2006 update. *Nucleic Acids Res*, 34:D411–D414, Jan 2006.

[191] B. Modrek and C. Lee. A genomic view of alternative splicing. *Nat Genet*, 30:13–19, Jan 2002.

[192] R. Mott, J. Schultz, P. Bork, and C. P. Ponting. Predicting protein cellular localization using a domain projection method. *Genome Res*, 12(8):1168–74, Aug 2002.

[193] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A. N. Nikolskaya, S. Orchard, M. Pagni, C. P. Ponting, E. Quevillon, J. Selengut, C. J. A. Sigrist, V. Silventoinen, D. J. Studholme, R. Vaughan, and C. H. Wu. Interpro, progress and status in 2005. *Nucleic Acids Res*, 33:D201–D205, Jan 2005.

[194] M. Nakao, R. A. Barrero, Y. Mukai, C. Motono, M. Suwa, and K. Nakai. Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals. *Nucleic Acids Res*, 33:2355–2363, Apr 2005.

[195] P. Nicolaidis and M. B. Petersen. Origin and mechanisms of nondisjunction in human autosomal trisomies. *Hum Reprod*, 13:313–319, Feb 1998.

[196] C. Nusbaum, M. C. Zody, M. L. Borowsky, M. Kamal, C. D. Kodira, T. D. Taylor, C. A. Whittaker, J. L. Chang, C. A. Cuomo, K. Dewar, M. G. FitzGerald, X. Yang, A. Abouelleil, N. R. Allen, S. Anderson, T. Bloom, B. Bugalter, J. Butler, A. Cook, D. DeCaprio, R. Engels, M. Garber, A. Gnirke, N. Hafez, J. L. Hall, C. H. Norman, T. Itoh, D. B. Jaffe, Y. Kuroki, J. Lehoczky, A. Lui, P. Macdonald, E. Mauceli, T. S. Mikkelsen, J. W. Naylor, R. Nicol, C. Nguyen, H. Noguchi, S. B. O'Leary, K. O'Neill, B. Piqani, C. L. Smith, J. A. Talamas, K. Topham, Y. Totoki, A. Toyoda, H. M. Wain, S. K. Young, Q. Zeng, A. R. Zimmer, A. Fujiyama, M. Hattori, B. W. Birren, Y. Sakaki, and E. S. Lander. Dna sequence and analysis of human chromosome 18. *Nature*, 437:551–555, Sep 2005.

[197] OBO. OBO: open biomedical ontologies, 2006. [Online; accessed 27-July-2006].

[198] OBO. MEDLINE, 2007. [Online; accessed 04-March-2007].

[199] K. P. O'Brien, M. Remm, and E. L. L. Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33:D476–D480, Jan 2005.

[200] T. Ogata and N. Matsuo. Comparison of adult height between patients with xx and xy gonadal dysgenesis: support for a y specific growth gene(s). *J Med Genet*, 29:539–541, Aug 1992.

[201] S. Orchard, H. Hermjakob, and R. Apweiler. The proteomics standards initiative. *Proteomics*, 3(7):1374–6, Jul 2003.

[202] S. Orchard, L. Montecchi-Palazzi, H. Hermjakob, and R. Apweiler. The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments. *Pac Symp Biocomput*, pages 186–96, 2005.

[203] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner. Predicting disease genes using protein-protein interactions. *J Med Genet*, 43:691–698, Aug 2006.

[204] C. A. C. Ottenheijm, L. M. A. Heunks, T. Hafmans, P. F. M. van der Ven, C. Benoist, H. Zhou, S. Labeit, H. L. Granzier, and P. N. R. Dekhuijzen. Titin and diaphragm dysfunction in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, 173:527–534, Mar 2006.

[205] A. Pandey and M. Mann. Proteomics to study genes and genomes. *Nature*, 405:837–846, Jun 2000.

[206] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia. Correlated mutations contain information about protein-protein interaction. *J Mol Biol*, 271(4):511–23, Aug 29 1997.

[207] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9):609–14, Sep 2001.

[208] F. Pazos and A. Valencia. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, 47(2):219–27, May 1 2002.

[209] A. G. Pedersen and H. Nielsen. Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *Proc Int Conf Intell Syst Mol Biol*, 5:226–33, 1997.

[210] C. Perez-Iratxeta, P. Bork, and M. A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nat Genet*, 31:316–319, Jul 2002.

[211] C. Perez-Iratxeta, M. Wjst, P. Bork, and M. A. Andrade. G2d: a tool for mining genes associated with disease. *BMC Genet*, 6:45, Aug 2005.

[212] N. Polavarapu, S. B. Navathe, R. Ramnarayanan, A. U. Haque, S. Sahay, and Y. Liu. Investigation into biomedical literature classification using support vector machines. *Proc IEEE Comput Syst Bioinform Conf*, pages 366–374, 2005.

[213] N. Przulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–15, Dec 12 2004.

[214] H. Qin, H. S. H. Lu, W. B. Wu, and W.-H. Li. Evolution of the yeast protein interaction network. *Proc Natl Acad Sci U S A*, 100(22):12820–4, Oct 28 2003.

[215] D. E. Quelle, F. Zindy, R. A. Ashmun, and C. J. Sherr. Alternative reading frames of the ink4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell*, 83:993–1000, Dec 1995.

[216] C. V. Rao and A. P. Arkin. Control motifs for intracellular regulatory networks. *Annu Rev Biomed Eng*, 3:391–419, 2001.

[217] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 67:026112, Feb 2003.

[218] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. Global variation in copy number in the human genome. *Nature*, 444:444–454, Nov 2006.

[219] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J.

Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of dna binding proteins. *Science*, 290:2306–2309, Dec 2000.

[220] D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 23(8):951–9., Aug 2005.

[221] R. L. Rivest. The MD5 message-digest algorithm. Internet Request for Comments, Apr. 1992. RFC 1321.

[222] R. J. Robbins. Bioinformatics: essential infrastructure for global biology. *J Comput Biol*, 3(3):465–78, Fall 1996.

[223] A. Robinson, B. G. Bender, and M. G. Linden. Summary of clinical findings in children and young adults with sex chromosome anomalies. *Birth Defects Orig Artic Ser*, 26:225–228, 1990.

[224] M. Roy, Q. Xu, and C. Lee. Evidence that public database records for many cancer-associated genes reflect a splice form found in tumors and lack normal splice forms. *Nucleic Acids Res*, 33:5026–5033, Sep 2005.

[225] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, Oct 2005.

[226] Rusch, Halpern, Sutton, Heidelberg, Williamson, Yooseph, Wu, Eisen, Hoffman, Remington, Beeson, Tran, Smith, Baden-Tillson, Stewart, Thorpe, Freeman, Andrews-Pfannkoch, Venter, Li, Kravitz, Utterback, Rogers, Falcón, Souza, Bonilla-Rosso, Eguiarte, Karl, Sathyendranath, Platt, Bermingham, Gallardo, Tamayo-Castillo, Ferrari, Strausberg, Nealson, Friedman, and Frazier. The sorcerer ii global ocean sampling expedition: Northwest atlantic through eastern tropical pacific. *PLoS Biol*, 5:e77, Mar 2007.

[227] R. Saito, H. Suzuki, and Y. Hayashizaki. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Res*, 30:1163–1168, Mar 2002.

[228] R. Saito, H. Suzuki, and Y. Hayashizaki. Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics*, 19:756–763, Apr 2003.

[229] G. Salton. Introduction to modern information retrieval. 1983.

[230] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res*, 32:D449–D451, Jan 2004.

[231] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, Oct 1995.

[232] B. J. A. Schijvenaars, B. Mons, M. Weeber, M. J. Schuemie, E. M. van Mulligen, H. M. Wain, and J. A. Kors. Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics*, 6:149, Jun 2005.

[233] A. Schlessinger and B. Rost. Protein flexibility and rigidity predicted from sequence. *Proteins*, 61(1):115–126., Aug 3 2005.

[234] G. R. Screaton, X. N. Xu, A. L. Olsen, A. E. Cowper, R. Tan, A. J. McMichael, and J. I. Bell. Lard: a new lymphoid-specific death domain containing receptor regulated by alternative pre-mrna splicing. *Proc Natl Acad Sci U S A*, 94:4615–4619, Apr 1997.

[235] M. Sekine, H. Nagata, S. Tsuji, Y. Hirai, S. Fujimoto, M. Hatae, I. Kobayashi, T. Fujii, I. Nagata, K. Ushijima, K. Obata, M. Suzuki, M. Yoshinaga, N. Umesaki, S. Satoh, T. Enomoto, S. Motoyama, and K. Tanaka. Localization of a novel susceptibility gene for familial ovarian cancer to chromosome 3p22-p25. *Hum Mol Genet*, 10:1421–1429, Jun 2001.

[236] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31:64–68, May 2002.

[237] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[238] N. R. Smalheiser and D. R. Swanson. Using arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed*, 57:149–153, Nov 1998.

[239] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome Biol*, 6:R46, Apr 2005.

[240] C. W. Smith and J. Valcárcel. Alternative pre-mrna splicing: the logic of combinatorial control. *Trends Biochem Sci*, 25:381–388, Aug 2000.

[241] G. R. Smith and M. E. J. Sternberg. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol*, 12(1):28–35, Feb 2002.

[242] M. Smith, V. Kunin, L. Goldovsky, A. J. Enright, and C. A. Ouzounis. Magicmatch–cross-referencing sequence identifiers across databases. *Bioinformatics*, 21(16):3429–30, Aug 15 2005.

[243] M. M. Sohocki, L. S. Sullivan, H. A. Mintz-Hittner, D. Birch, J. R. Heckenlively, C. L. Freund, R. R. McInnes, and S. P. Daiger. A range of clinical phenotypes associated with mutations in crx, a photoreceptor transcription-factor gene. *Am J Hum Genet*, 63:1307–1315, Nov 1998.

[244] E. L. Sonnhammer, G. V. Heijne, and A. Krogh. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6:175–82, 1998.

[245] R. S. Spielman, L. A. Bastone, J. T. Burdick, M. Morley, W. J. Ewens, and V. G. Cheung. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet*, 39:226–231, Feb 2007.

[246] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, 100:12123–12128, Oct 2003.

[247] E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–92, Aug 24 2001.

[248] L. Stein. Creating a bioinformatics nation. *Nature*, 417:119–120, May 2002.

[249] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mint-

zlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122:957–968, Sep 2005.

[250] R. D. Stevens, A. J. Robinson, and C. A. Goble. mygrid: personalised bioinformatics on the information grid. *Bioinformatics*, 19 Suppl 1:i302–i304, 2003.

[251] C. S. Stipp, T. V. Kolesnikova, and M. E. Hemler. Functional domains in tetraspanin proteins. *Trends Biochem Sci*, 28:106–112, Feb 2003.

[252] M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci U S A*, 102:4221–4224, Mar 2005.

[253] J. Sua, K. Richtera, C. Zhangb, Q. Guc, and L. Li. *Molecular Immunology*, 2006.

[254] D. R. Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med*, 30:7–18, 1986.

[255] D. R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med*, 31:526–557, 1988.

[256] D. R. Swanson. Atrial fibrillation in athletes: implicit literature-based connections suggest that overtraining and subsequent inflammation may be a contributory mechanism. *Med Hypotheses*, 66:1085–1092, Feb 2006.

[257] T. D. Taylor, H. Noguchi, Y. Totoki, A. Toyoda, Y. Kuroki, K. Dewar, C. Lloyd, T. Itoh, T. Takeda, D.-W. Kim, X. She, K. F. Barlow, T. Bloom, E. Bruford, J. L. Chang, C. A. Cuomo, E. Eichler, M. G. FitzGerald, D. B. Jaffe, K. LaButti, R. Nicol, H.-S. Park, C. Seaman, C. Sougnez, X. Yang, A. R. Zimmer, M. C. Zody, B. W. Birren, C. Nusbaum, A. Fujiyama, M. Hattori, J. Rogers, E. S. Lander, and Y. Sakaki. Human chromosome 11 dna sequence and analysis including novel gene identification. *Nature*, 440:497–500, Mar 2006.

[258] S. A. Teichmann and R. A. Veitia. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics*, 167:2121–2125, Aug 2004.

[259] L. H. Thompson. Unraveling the fanconi anemia-dna repair connection. *Nat Genet*, 37:921–922, Sep 2005.

[260] L. Tsyba, I. Skrypkina, A. Rynditch, O. Nikolaienko, G. Ferenets, A. Fortna, and K. Gardiner. Alternative splicing of mammalian intersectin 1: domain associations and tissue specificities. *Genomics*, 84:106–113, Jul 2004.

[261] A. L. Turinsky, A. C. Ah-Seng, P. M. K. Gordon, J. N. Stromer, M. L. Taschuk, E. W. Xu, and C. W. Sensen. Bioinformatics visualization and integration with open standards: the bluejay genomic browser. *In Silico Biol*, 5:187–198, 2005.

[262] F. S. Turner, D. R. Clutterbuck, and C. A. M. Semple. Pocus: mining genomic sequence annotation to predict disease genes. *Genome Biol*, 4:R75, Oct 2003.

[263] E. Turpin, B. Dalle, A. de Roquancourt, L. F. Plassa, M. Marty, A. Janin, Y. Beuzard, and H. de Thé. Stress-induced aberrant splicing of tsg101: association to high tumor grade and p53 status in breast cancers. *Oncogene*, 18:7834–7837, Dec 1999.

[264] M. Tyers and M. Mann. From genomics to proteomics. *Nature*, 422:193–197, Mar 2003.

[265] UDDI. UDDI: Universal description, discovery and integration protocol, 2006. [Online; accessed 27-July-2006].

[266] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, Feb 2000.

[267] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, 12(3):368–73, Jun 2002.

[268] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 14:535–542, May 2006.

[269] M. A. van Driel, K. Cuelenaere, P. P. C. W. Kemmeren, J. A. M. Leunissen, and H. G. Brunner. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet*, 11:57–63, Jan 2003.

[270] M. A. van Driel, K. Cuelenaere, P. P. C. W. Kemmeren, J. A. M. Leunissen, H. G. Brunner, and G. Vriend. Geneseeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res*, 33:W758–W761, Jul 2005.

[271] C. G. Van Horn, J. M. Caviglia, L. O. Li, S. Wang, D. A. Granger, and R. A. Coleman. Characterization of recombinant long-chain rat acyl-coa synthetase isoforms 3 and 6: identification of a novel variant of isoform 6. *Biochemistry*, 44:1635–1642, Feb 2005.

[272] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

[273] F. M. Vaz, R. H. Houtkooper, F. Valianpour, P. G. Barth, and R. J. A. Wanders. Only one splice variant of the human taz gene encodes a functional protein with a role in cardiolipin metabolism. *J Biol Chem*, 278:43089–43094, Oct 2003.

[274] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner,

S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291:1304–1351, Feb 2001.

[275] A. Vespignani. Evolution thinks modular. *Nat Genet*, 35:118–119, Oct 2003.

[276] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, May 2002.

[277] W3C. W3C definition and mission statement, 2006. [Online; accessed 24-July-2006].

[278] A. J. Walhout, R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg, and M. Vidal. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287:116–122, Jan 2000.

[279] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, Jun 1998.

[280] C. A. Wells, A. M. Chalk, A. Forrest, D. Taylor, N. Waddell, K. Schroder, S. R. Himes, G. Faulkner, S. Lo, T. Kasukawa, H. Kawaji, C. Kai, J. Kawai, S. Katayama, P. Carninci, Y. Hayashizaki, D. A. Hume, and S. M. Grimmond. Alternate transcription of the toll-like receptor signaling cascade. *Genome Biol*, 7:R10, Feb 2006.

[281] S. Wernicke. A faster algorithm for detecting network motifs. In *Proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI '05)*, volume 3692 of *LNBI*, pages 165–177. Springer, 2005.

[282] Wikipedia. Computational complexity theory — wikipedia, the free encyclopedia, 2006. [Online; accessed 6-July-2006].

[283] Wikipedia. Moore's law — wikipedia, the free encyclopedia, 2006. [Online; accessed 5-July-2006].

[284] Wikipedia. Uniform resource identifier — wikipedia, the free encyclopedia, 2006. [Online; accessed 5-July-2006].

[285] M. D. Wilkinson and M. Links. Biomoby: an open source biological web services proposal. *Brief Bioinform*, 3:331–341, Dec 2002.

[286] J. Wojcik and V. Schächter. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17 Suppl 1:S296–305, 2001.

[287] W. M. Wojtowicz, J. J. Flanagan, S. S. Millard, S. L. Zipursky, and J. C. Clemens. Alternative splicing of drosophila dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, 118:619–633, Sep 2004.

[288] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek. The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Res*, 34:D187–D191, Jan 2006.

[289] S. Wuchty, Z. N. Oltvai, and A.-L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet*, 35:176–179, Oct 2003.

[290] Y. Xing and C. Lee. Alternative splicing and rna selection pressure–evolutionary consequences for eukaryotic genomes. *Nat Rev Genet*, 7:499–509, Jul 2006.

[291] Q. Xu, B. Modrek, and C. Lee. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*, 30:3754–3766, Sep 2002.

[292] B. K. Yeh, M. Igarashi, A. V. Eliseenkova, A. N. Plotnikov, I. Sher, D. Ron, S. A. Aaronson, and M. Mohammadi. Structural basis by which alternative splicing confers specificity in fibroblast growth factor receptors. *Proc Natl Acad Sci U S A*, 100:2266–2271, Mar 2003.

[293] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4:928–942, Apr 2004.

[294] G. U. Yule. A mathematical theory of evolution, based on the conclusions of dr jc willis, frs. *Phil Trans Roy Soc Lond Series B*, 213:21–87, 1925.

[295] A. Yuryev, Z. Mulyukov, E. Kotelnikova, S. Maslov, S. Egorov, A. Nikitin, N. Daraselia, and I. Mazo. Automatic pathway building in biological association networks. *BMC Bioinformatics*, 7:171, Mar 2006.

[296] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. Mint: a molecular interaction database. *FEBS Lett*, 513:135–140, Feb 2002.

[297] C. Zhang, H.-R. Li, J.-B. Fan, J. Wang-Rodriguez, T. Downs, X.-D. Fu, and M. Q. Zhang. Profiling alternatively spliced mrna isoforms for prostate cancer classification. *BMC Bioinformatics*, 7:202, Apr 2006.

[298] T. Zhang, P. Haws, and Q. Wu. Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation. *Genome Res*, 14:79–89, Jan 2004.

[299] X.-H. Zhou, D. K. McClish, and N. A. Obuchowski. *Statistical Methods in Diagnostic Medicine*. Wiley, 2002.

[300] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein, and M. Snyder. Global analysis of protein activities using proteome chips. *Science*, 293(5537):2101–5, Sep 14 2001.